Universidad Autónoma de Madrid

Doctoral Thesis

Geocomputation methods for spatial economic analysis

Author:
Andres Marcelo Vallone

Supervisor:
Dr. Coro Chasco

A thesis submitted in fulfillment of the requirements
for the degree of Doctor of Philosophy in Economics and Business
in the
Departamento de Economía Aplicada
of the
Facultad de Ciencias Económicas y Empresariales

2019

Thesis Committe

Geoffrey Hewing
University of Illinois Urbana-Champaign, U.S.A

Román Mínguez Salido
University of Castilla-La Mancha Cuenca, Spain

Luis Collado Cueto
Autonomous University of Madrid Madrid, Spain

Alternate Committe

Sofia Garcia Games
Autonomous University of Madrid Madrid, Spain

Sofia Ruiz Campos
EAE Business School Campus Madrid Madrid, Spain

Externals Evaluators

Patricio Aroca
Adolfo Ibáñez University Viña del Mar, Chile

Ludo Peters
Hasselt University Hasselt, Belgium

"*Computing is not about computers anymore. It is about living.*"

Nicholas Negroponte

UNIVERSIDAD AUTÓNOMA DE MADRID

Abstract

Facultad de Ciencias Económicas y Empresariales

Departamento de Economía Aplicada

Doctor of Philosophy in Economics and Business

Geocomputation methods for spatial economic analysis

by Andres Marcelo Vallone

Geocomputation is a new scientific paradigm that uses computational techniques to analyze spatial phenomena. Spatial economics and regional science quickly adopted geocomputation techniques to study the complex structures of urban and regional systems. This thesis contributes to the use of geocomputation in spatial economic analysis through construction and application of a new set of algorithms and functions in the R programming language to deal with spatial economic data. First, we created the 'DataSpa' package, which collects data at low geographical levels to generate socio-economic information for Spanish municipalities using URL parsing, PDF extraction and web scraping. Second, based on a search and replace algorithm, we built the 'msp' package to harmonize data with accuracy problems such as spelling errors, acronym abbreviations and names listed differently. This methodology enables study of the patenting activity and research collaboration in Chile between 1989-2013. We also adapted classical spatial autocorrelation methods to visualize and explore the existence of productivity spillovers among the network's members. Finally, we created 'estdaR' to improve knowledge of Chile's urban system by evaluating the influence of spatial proximity among human settlements on the evolution of cities. The package contains new tools for exploratory spatio-temporal data analysis that are very useful for detecting spatial differences in time trends. All R codes used in computation and the packages themselves are considered as research results and are freely available to other researchers in a Github repository.

Keywrods: Geocomputations, R programing, Urban Studies, Patent analysis Chile, Spain
JEL codes: C81, C88, C63, R12

UNIVERSIDAD AUTÓNOMA DE MADRID

Resumen

Facultad de Ciencias Económicas y Empresariales

Departamento de Economía Aplicada

Doctorado en Economía y Empresa

Geocomputation methods for spatial economic analysis

por Andres Marcelo Vallone

La Geocomputatión es un nuevo paradigma científico que utiliza métodos computacionales para analizar fenómenos espaciales. La economía espacial y la ciencia regional adoptaron rápidamente las técnicas de la geocomputación para estudiar las estructuras complejas de los sistemas urbanos y regionales. Esta Tesis constituye una contribución al campo de la geocomputación a través de la construcción y la aplicación al análisis económico espacial de un nuevo conjunto de algoritmos y funciones programadas en lenguaje R. En primer lugar, utilizando técnicas de análisis sintáctico de las URL, de extracción de textos en formato PDF y de "web scraping", hemos desarrollado el paquete "DataSpa" que recopila información procedente de Internet necesaria para generar indicadores socioeconómicos para los municipios españoles. En segundo lugar, utilizando un algoritmo de búsqueda y remplazo se genera el paquete "msp" que permite arreglar textos con imprecisiones y errores de escritura en los acrónimos y nombres propios. De esta forma, fue posible estudiar las relaciones de colaboración empresarial y la actividad de I+D de las empresas chilenas, en el período 1989-2013, a través de las relaciones en materia de patentes. Adicionalmente, hemos adaptado métodos clásicos de autocorrelación espacial a este ámbito para explorar y visualizar la existencia de efectos de contagio en la productividad de la red de colaboración en la actividad de I+D entre las empresas. Finalmente, para mejorar el conocimiento del sistema urbano chileno, hemos evaluado la influencia que la proximidad espacial entre ciudades tiene en la evolución de su tamaño poblacional, a través del paquete "estdaR", que contiene funciones para el análisis exploratorio de datos espacio-temporales que permiten para analizar diferencias espaciales en las tendencias temporales. Todos los códigos de R usados y los paquetes son considerados, en sí mismos, un resultado de la investigación y están libremente disponibles en un repositorio en Github

Palabras clave:geocomputación, programa R, estudios urbanos, análisis de patentes, Chile, España

JEL codes: C81, C88, C63, R12

Agradecimientos

Quiero agradecer en primer lugar a Dios, por darme la oportunidad de vivir esta experiencia y por ayudarme a cumplir este objetivo que surgió al instante de terminar mi Licenciatura en San Juan cuando le dije a mi nonna que sería que en algún momento sería un doctor, tal como ella me sugirió al momento de elegir la carrera universitaria haciendo referencia a seguir la carrera de medicina. El camino hasta aquí fue largo y lleno de experiencias. La primera parada fue en la aquel entonces Escuela de Ingeniería Comercial y actual Escuela de Ciencias Empresariales de la Universidad Católica del Norte en Coquimbo Chile, donde concluí los estudios de Master en Administración y conocí un maravillo grupo de personas que ahora son mis colegas a quienes estoy inmensamente agradecido por aceptar llevar a cabo mis labores mientras yo terminaba mis estudios doctorales. En está etapa di mis primeros pasos hacia la investigación de la mano de Sergio Zuñiga y Ramón Lacayo.

La segunda parada de este camino es la Facultad de Administración y Economía de la misma universidad en la ciudad de Antofagasta, allí tuve contacto con la Ciencia Regional que determino mis deseos de desempeñarme en esta área, allí conocí a grandes profesores y personas a las cuales estoy rotundamente agradecido como Nicolas Garrido, quien me introdujo al mundo de R y a las cadenas de Markow, Patricio Aroca que me enseño econometría espacial, usando el manual de quien sería posteriormente mi directora de tesis doctoral, Miguel Atienza quien me enseño las primeras nociones de economía urbana y localización y Marcelo Lufin quien me enseño el uso de las redes sociales, todos ellos han colaborado en está tesis doctoral al enseñarme los fundamentos y un gran arsenal de herramientas muy necesarias para realizar investigaciones de calidad. La tercer parada nos lleva de regreso a la Escuela de Ciencias Empresariales en Coquimbo donde actualmente me desempeño como profesor, esta etapa afianzo mis deseos de realizar el doctorado y quiero agradecer el apoyo de la Universidad Católica del Norte a través de su beca y mis compañeros Andres Araya, Paulina Gutierrez, Karla Soria, Pablo Pinto, Sergio Zuñiga, Cristian Morales, Rodrigo Sfeir, Carolina Fuentes quienes han estado constantemente pendientes y dispuestos para ayudar durante todo el periodo.

La última parada de este viaje es Madrid, en el Departamento de Economía Aplicada donde he concluido esta tesis. Quiero agradecer a mi directora Coro Chasco, gracias por su amistad, paciencia y constante esmero por lograr de mi un mejor profesional, guiándome e insistiendo constantemente en abrir mis horizontes impulsándome a ir Cartagena, Tarragona, Barcelona Milán, Minneapolis y Dijon a presentar mis trabajos. Agradezco inmensamente su confianza y la libertad creativa que me ha dado para realizar esta tesis. Quiero agradecer a mis compañeros del grupo de investigación de Economía Regional y Espacial (ECONRES), en especial a Ludo Peeters, Beatriz Sánchez, Sofía Ruiz Campos, Enrique Marinao y Julie Le Gallo, por su

apoyo constante y su amistad. Asimismo, también quiero agradecer a los investigadores del Instituto Lawrence R. Klein, en especial, a su Director, José Vicéns y al coordinador del Área de Economía Espacial Microterritorial, Pedro Chasco, que me hayan permitido colaborar en algunos de sus proyectos, pues han iluminado parte de mi Tesis Doctoral. Además, vaya mi reconocimiento también para el Departamento de Economía Aplicada al que pertenezco, en calidad de Personal Investigador en Formación (PIF), de un modo especial a su Directora, Milagros Dones, a la profesora Sofía García y a la gestora de la UDI de Econometría e Informática, Aniana Peña, por haberme acogido y apoyado en los muchos trámites que he tenido que realizar. De un modo especial, quisiera también reconocer las sugerencias y aportaciones formuladas por los profesores que formaron parte del tribunal de pre-defensa de mi Tesis Doctoral: Julián Pérez (Departamento de Economía Aplicada), Joan Crespo y Carlos Fernández (Departamento de Economía Cuantitativa), que formaron parte del tribunal de pre-defensa doctoral de mi Tesis Doctoral, así como al profesor Francisco Vázquez (Departamento de Economía Cuantitativa), que me acogió como alumno en sus clases de Sistemas Dinámicos. Quiero agradecer también a los miembros del tribunal de mi Tesis Doctoral, los profesores Geoffrey Hewings (Universidad de Illinois en Urbana-Champaign, USA), Román Mínguez (Universidad de Castilla-La Mancha) y Luis Collado (Departamento de Estructura Económica y Economía del Desarrollo de la Universidad Autónoma de Madrid) por aceptar revisar mi obra y formar parte de este comité de defensa doctoral.

Sin dudas este largo viaje no hubiese sido posible sin el apoyo incondicional de mi esposa. Gracias, Camila, por abandonar todo y seguirme en esta aventura. Gracias por tu paciencia, sobre todo en las etapas de lejanía durante los viajes y las estancias de investigación lejos de casa. Sin tu apoyo esto no hubiese sido posible. También agradecer a mi compañera de aventuras: gracias hija, por soportar tan valientemente la lejanía y dar tanto amor, además de elegir correctamente todos los colores de mis presentaciones. Debo admitir que son la fuente de mi inspiración y cualquier complicación es asumible a su lado.

Un camino largo permite generar un gran conjunto de amistades, que en gran medida cuando estamos fuera del lugar donde se encuentra nuestra propia familia, comienzan a constituir una nueva familia. Esta familia por elección es amplia, una parte se encuentra en Chile. A ellos agradezco que hayan entendido y aceptado la distancia, además de su constante preocupación por nuestro bienestar. A mis amigos Marcelo Olivares, Cristian Morales y Rodrigo Sfeir, gracias por estar ahí en los momentos difíciles. Quiero agradecer a los miembros en España de esta familia: a Beatriz y José (¡Buen Camino amigos!), a Elvio, Andrea, Ailen, Emiliano y Juan, a Sofía y Tony y a Jesús y Amparo. Gracias por su apoyo y amistad que han hecho que la estancia durante el desarrollo de esta obra sea maravillosa, logrando que nos sintamos como en nuestro propio hogar.

Por último, agradecer a mis padres y hermanos, tanto sanguíneos como políticos, que han

aceptado estoicamente la distancia y han estado siempre presentes en las etapas más difíciles de este viaje.

Acknowledgements

I would like to acknowledge to Julie LeGallo, Kassoum Ayouba and the rest of the researchers of the Centre d'Économie et de Sociologie Appliquées à l'Agriculture et aux Espaces Ruraux, AgroSup-Dijon for helping me during my stay in Dijon. Thank you for contributing with your advice in the last chapter of this Dissertation. My experience at the Center was excellent and I would like to have the opportunity to go back again.

Although previously cited, I also like to thank Geoffrey Hewing in his own language for showing his interest in my work and accepting crossing the Ocean to be part of the Thesis Committee.

Contents

List of Figures

List of Tables

List of Algorithms

To Camila and Francesca

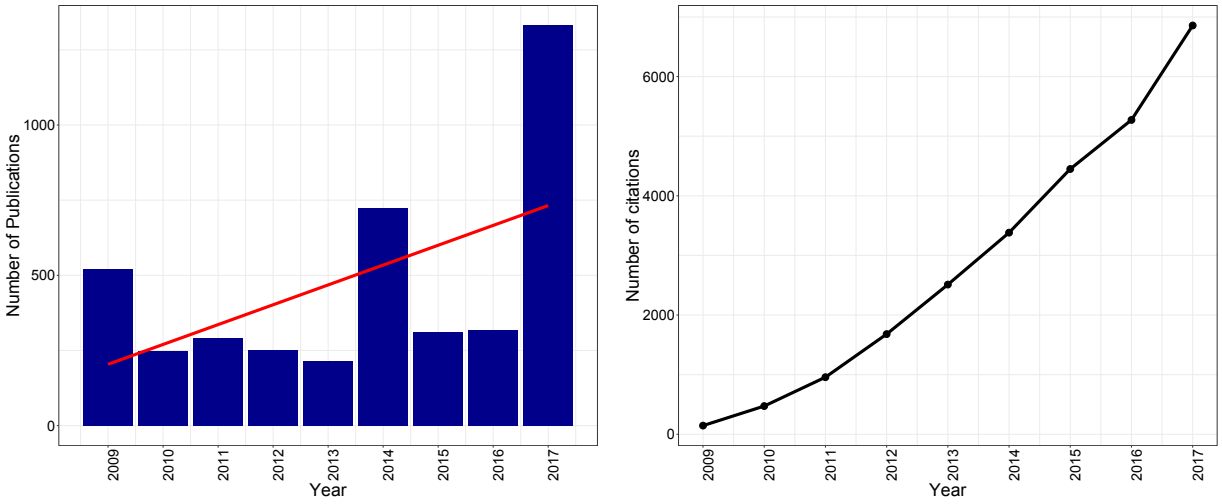
1 Introduction

Use of the term "geocomputation" began in 1996 with the first international conference on 'Geocomputation', hosted by the School of Geography of the University of Leeds, launching a new research agenda in geographical analysis and modeling (Openshaw & Abrahart, 1996). The increase in large-scale computation caused by this technological advance has increased the importance of the term geocomputation (Batty, 2017). A quick search of the topic 'geocomputation' in the Dimensions database[1] shows evidence of an increasing growth of this field. Between 2009 and 2017, 4212 publications were related to geocomputation topics. Figure 1.1(a) shows the evolution of the number of publications per year, which follows a positive linear trend (plotted in red). In terms of impact, the topic has 6.1 citations per publication in average and a total number of 25728 citations (Figure 1.1(b)).

Several definitions of geocomputation have been proposed. Rees and Turton, 1998 define it as "*the process of applying computing technology to geographical problems". For Couclelis, 1998, "geocomputation just means the universe of computational techniques applicable to spatial problems*". Openshaw, 2014 considers geocomputation as a new paradigm "*concerned with the application of a computational science paradigm to study all manner of geophenomena including both physical and human systems*" (Cheng, Haworth, & Manley, 2012) also provide a simple and useful definition of geocomputation as "*[t]he art and science of solving complex spatial problems with computers*". Thus, geocomputation is not simply about applying computational methods to explore geographical concepts. It offers an extensive toolkit for examination and identification of new perspectives on spatial processes (Cheng et al., 2012).

Despite absence of consensus on the definition of geocomputation, the framework clearly has four main dimensions: (1) computer architecture and design; (2) search, classification, prediction and modelling; (3) knowledge discovery and (4) visualization (Cao, Ge, & Wang, 2014; Cheng et al., 2012; Fischer & Leung, 2001; Gahegan, 1999; Openshaw & Abrahart, 1996). The first, computer architecture and design, addresses the improvements in computing performance and the development of new architectures. The second—search, classification, prediction and modelling—is related to progress in pattern recognition, classification and function approximation tools. Third, knowledge discovery advances the mechanisms by

[1] https://www.dimensions.ai/



(a) Publication Numbers (b) Citation Number

FIGURE 1.1: Growth of geocumputation topics
Source: self-elaboration

which knowledge can be distilled from large datasets as a result of data mining and knowledge discovery tools. Finally, visualization addresses advances in visualization as a means of data exploration, providing new tools and approaches that give insight into complex and multi-dimensional datasets (Gahegan, 1999). These dimensions are not mutually exclusive; they are closely interrelated. One could even argue that the areas of search, classification, prediction and modelling area and of knowledge discovery differ only in the algorithm used (Cheng et al., 2012).

Additionally, geocomputation associated with topics like spatial big data (Li et al., 2016; Thakuriah, Tilahun, & Zellner, 2017), machine learning (Marjanović, Bajat, Abolmasov, & Kovačević, 2018; Pradhan, 2013; Rogan et al., 2008), parallel computing (Guan, Hu, Liu, & Yun, 2018; Tang, Feng, Deng, Jia, & Zuo, 2018), agent-based models (Patel, Crooks, & Koizumi, 2018; Ramírez, Leger, & Vallone, 2014; Wang, 2018), software development and software applications (Bivand, 2014; Guan et al., 2018; Rey, 2015), among other methods.

Since geocomputation's focus is spatial analysis (Fischer, 2006), the field of spatial economics—and especially the regional science—has adopted it to study the complex structures of urban and regional systems (Thill & Dragicevic, 2018). Methodologies traditional to the geocomputation framework, such as agent-based simulation and microsimulation, have been effective at interfacing bottom-up computational principles with the fundamental theories of behavioral, social and economic sciences to advance understanding of the complex organizations of regional systems (Thill & Dragicevic, 2018). Table 1.1 shows the results of a Web of Science search on those publications indexed in the research areas of Economics and Social

Sciences during the period of 2000 and 2018, which are connected with three geocomputational topics 'agent based', 'big data' and 'machine learning'). In a second phase, we filtered out those articles and book chapters specifically associated with urban and regional science, geography and geoscience, in order to calculate their corresponding share in the total sum. These three topics have experienced a growth in terms of publications over the last 5 years. The share of Urban and Regional Science is over 9% during the entire period demonstrating the existence of a strong link between the fields of geocomputation and spatial economics. In fact, some authors are talking about a new 'GIScience paradigm' (Thill & Dragicevic, 2018; Vaz, 2016, 2018)

TABLE 1.1: Number of publications of geocomputational topics in economics.

Topic	Area	2013	2014	2015	2016	2017	Full period
Agent Based Models	Total Economics	68	79	96	116	168	1012
	Urban/Regional	7	8	17	11	22	128
	Participation	10%	10%	18%	9%	13%	13%
Big Data	Total Economic	26	28	67	73	99	418
	Urban/Regional	11	12	36	34	38	172
	Participation	42%	43%	54%	47%	38%	41%
Machine Learning	Total Economic	3	3	4	12	36	69
	Urban/Regional	3	2	1	8	21	37
	Participation	100%	67%	25%	67%	58%	54%

Source: self-elaboration from the Web of Science

The dissertation aims to contribute to the use of geocomputation in spatial economic analysis through construction and application of a new set of algorithms and functions in the R programming language to analyze spatial economic data. The dissertation addresses three different topics in spatial economics that require development of new geocomputational tools for analysis to deal with particular problems associated with each one of them (figure 1.2). The dissertation thus makes a twofold contribution, to spatial economic analysis and to the geocomputational paradigm. It contributes to the topic of search, classification, prediction and modelling in geocomputation by developing two R packages, the "DataSpa" package (Vallone, Chasco, & Sanchez, 2017) to collect spatial information and the "msp" package (Vallone, 2018) to improve data quality. It contributes to knowledge discovery in geocomputation by developing new implementation of exploratory spatio-temporal data analysis tools in the "estdaR" R package (Vallone, Le Gallo, Chasco, & Ayouba, 2018). Finally, it contributes to geocomputational visualization by adapting classical spatial econometric techniques to network visualization to explore the existence of productivity spillovers between a network's members. The application of these tools in this Thesis to specific empirical cases for Spain and Chile provides a better knowledge of some socioeconomic aspects of these countries and makes in this way a substantial contribution to the spatial economic literature.

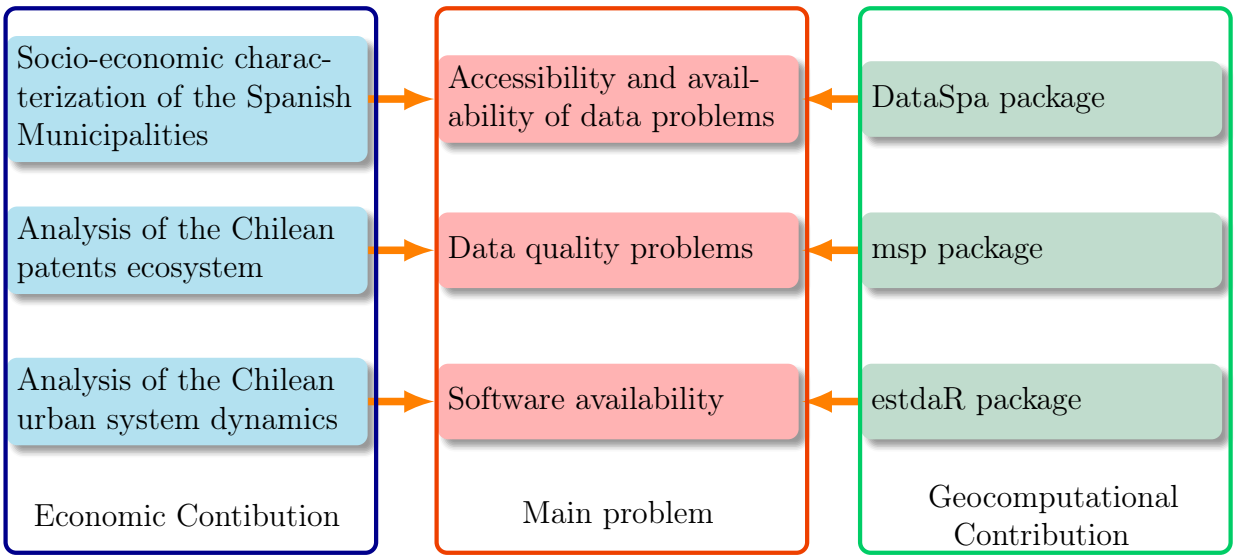


FIGURE 1.2: Dissertation Contributions
Source: self-elaboration

1.1 Research topics

This dissertation addresses a number of research topics spanning the typical process of economic analysis, from data collection, cleaning and visualization to statistical analysis. First, it considers problems of accessibility and availability commonly present in web-based data collection, depending on data category and country, since online information is often only provided via parochial and hard-to-use web interfaces (Foster, 2011). This is the case for some official and non-official databases of socioeconomic information published in Spain at finer spatial scales, to which access is restricted or even blocked. To overcome some of these problems, Chapter 3 of the dissertation presents various strategies based on URL parsing, PDF text extraction and web scraping to extract and organize several databases in Spain at the level of municipality. Each strategy consists of a set of functions constructed in the R package "DataSpa". Second, large databases usually present problems of information quality. Data accuracy can be affected by human error in database construction, such as spelling typos and name repetition. Chapter 4, a large database of patents in Chile, is cleaned by means of a search and replace algorithm "msp", built in R, which harmonizes the database information based on assignee names and avoiding misspellings and duplications.

Third, once the data are collected, cleaned and organized, it can be challenging to explore and visualize highly dimensional databases at finer resolutions (Brunsdon & Singleton, 2015). It is crucial, however, to find interesting underlying patterns, structures and relationships to provide reliable decision-support (Thill & Dragicevic, 2018). The dissertation addresses this problem by proposing new routines in R to explore the existence of productivity spillovers

into the Chilean patent co-affiliation network. Chapter 5 also proposes a collection of exploration tools for spatio-temporal data analysis, as applied to the study of Chilean urban system dynamics during the last century.

Finally, the dissertation considers the increasing importance of construction and publication of open source packages in spatial economic analysis. Researchers tend to develop scripts or functions to ensure reproducibility, currently a major issue for data analysts (Brunsdon, 2014). Sharing codes and functions related to similar topics boosts visibility and dissemination of a specific study. As explained previously, one goal of this dissertation is to develop open source tools for analysis of spatial economic databases. To this end, Chapter 5 presents a new R package, "estaR", which compiles statistical tools for spatio-temporal analysis of city evolution in Chile.

1.2 Relevance of the research

1.2.1 Chapter 3 (Paper 1)

Chapter 3 aims to solve problems related to data collection at low geographic levels for urban studies proposed in Spain. Official agencies provide data on the variables of population, workforce and vehicle fleet in Spain through their web portals–the National Statistics Office (INE)[2], Spanish Public Employment Service (SEPE)[3] and Spanish National Department of Traffic (DGT)[4], respectively. However, these data sets are difficult to access. The sophisticated organizational structure of the official agencies' web portals, absence of adequate APIs and/or lack of coding of the data make the downloading process arduous and highly vulnerable to human failure. To overcome these problems of accessibility, we use URL parsing methodology to analyse the URLs of the official agencies' web portals, build a set of functions to connect to the agency server, download the required information and manipulate it into R. Further, since the DGT information is available in PDF format, we apply a pdf extraction strategy to build a set of R functions that load and manipulate the vehicle fleet information into R. In Spain, firm data at micro level also present availability problems. The INE's Central Company Directory (DIRCE)[5] provides annual information on the number and distribution of companies and establishments, but only at meso- and macro-level in order to preserve statistical confidentiality. The Industrial Establishment Register (RIEI)[6],

[2]http://www.ine.es.
[3]http://www.sepe.es.
[4]http://www.dgt.es.
[5]http://www.ine.es/dynt3/inebase/es/index.htm?padre=51&dh=1
[6]http://www.minetad.gob.es/industria/RII/Paginas/Index.aspx

published by the Ministry of Energy, Tourism and Digital Agenda, contains a census of industrial establishments located in Spain at municipality level and above, but this census is incomplete for some regions and industrial sectors, only includes companies operating in the industrial sector and has some downloading restrictions. To overcome these problems and create a company and freelance database for Spain at municipality level and above, we drew on the web scraping technique to build a set of R functions to browse and extract online information provided freely by the private consulting firm Axesor[7] All of the functions form part of the "DataSpa" R package, available under GPL-2 license (Vallone et al., 2017).

1.2.2 Chapter 4 (Paper 2)

In Chapter 4 we focus in study patenting activity and research collaboration in Chile between 1989 and 2013 based on a patent dataset compiled from the Chilean National Institute of Industrial Property (INAPI). The aim this chapter is provide a quantitative overview of the patenting landscape in Chile during the period of 1989 – 2013 and using social network analysis (Scott, 2013) explore the country's patterns of research collaboration. Different from most industrialized economies, comprehensive databases are often incomplete, and researchers must collect data directly on a case-by-case basis. In the particular case of Chile, the raw data present discrepancies such as spelling errors, acronyms, abbreviations, and names listed differently. Different assignee names referring to the same individual or entity were also found. Some assignee names change over time, being necessary track it to modify it. Due to the raw data has more than 42.000 record, fix these data quality problems manually is an ineffective and very exposed to human failures task, for that reason based on a search and replace algorithm we build an interactive R function to solve them. This function is part of the packages "msp" also available under GPL-2 license (Vallone, 2018). We build a network of affiliations (Scott, 2013) where assignees or patent owners are actors and patents are events to analyze the Chilean patent ecosystem and its research collaboration network. In addition, we transform this bimodal affiliation network into a unimodal network and used its adjacency matrix as "vicinity" weights in the calculation of the Moran's I and global G autocorrelation tests to contrast the relationship between productivity and collaboration. To explore the existence of productive local cluster in the network, we depict it using the four quadrants of the Moran's scatterplot, classifying the assignees into four mutually and exclusive categories. Overall, our results show the lack of collaboration between industrial companies as well as science-industry in Chile, at least in terms of patent co-ownership. Evidence reveals an intensification of knowledge content out of domestic and foreign companies, but also a small contribution of universities to such knowledge. We also found that grant lags were long and variable. This variation is driven

[7] https://www.axesor.es.

by differences across technology fields, experience, residency, and type of assignees. All the R code used in computation are freely available to other researchers interested in using it.

1.2.3 Chapter 5 (Paper 3)

The aim of chapter 5 is focus on improving the knowledge of the Chilean urban system through a set of novel tools, which allows evaluating the influence of spatial proximity among human settlements on the evolution of the cities to detect regional differences in these spatio-temporal dynamics. We want to respond to the following questions: First, are certain urban processes homogeneous across the entire Chilean city system? Second, is the Duranton's hypothesis (Duranton, 2016) about the existence of stronger agglomeration effects in less-advanced countries applicable to Chile too? Third, to what extent will a city population grow faster or slower depending on its neighbor's growth speed?. To answer these questions, we use the census data over the period 1930-2002 and first we analyze the cross-sectional distribution of urban population by means of standard statistical analysis and nonparametric estimations of density functions for some particular years, as proposed by Quah, 1996and followed by many other authors (e.g. Xu and Zhu, 2009 and Xiufang et al., 2015). Second, the growth process is modeled as a first-order stationary Markov chain and the role of geographical space on the transition probabilities is evaluated with a set of methods based on a spatial version of the standard Markov chain (Le Gallo & Chasco, 2008). Third, we also perform an in-depth analysis to detect spatial regimes in the movement direction and ranking mobility of the Chilean urban distribution. The LISA Markov (Rey & Janikas, 2006) and directional LISA (Rey, Murray, & Anselin, 2011) approaches capture the co-evolution of a spatial unit with its respective neighbors identifying different spatial regimes in the ranking mobility of the urban distribution. We also we study the existence of spatial differences in the own growth pattern with the Global Indicator of Mobility Association (GIMA) (Rey, 2016). Finally, the Ranking decomposition (Rey, 2004) is a cohesion measure that detects synchronic rank movements among spatial regimes. All the calculus was performed using the "estdaR" R package (Vallone et al., 2018) available under GPL-2 license. The results show the existence of different regional dynamics, reflecting the existence of spatial heterogeneity in the Chilean urban system, there are also evidence of a clear and persistent pattern of agglomeration economies in the Chilean urban system. Finally, the probability for a city to grow increases with its neighbors' size, while large cities surrounded by smaller towns will not experience practically any change in population, hence, spatial proximity matters in the Chilean urban system.

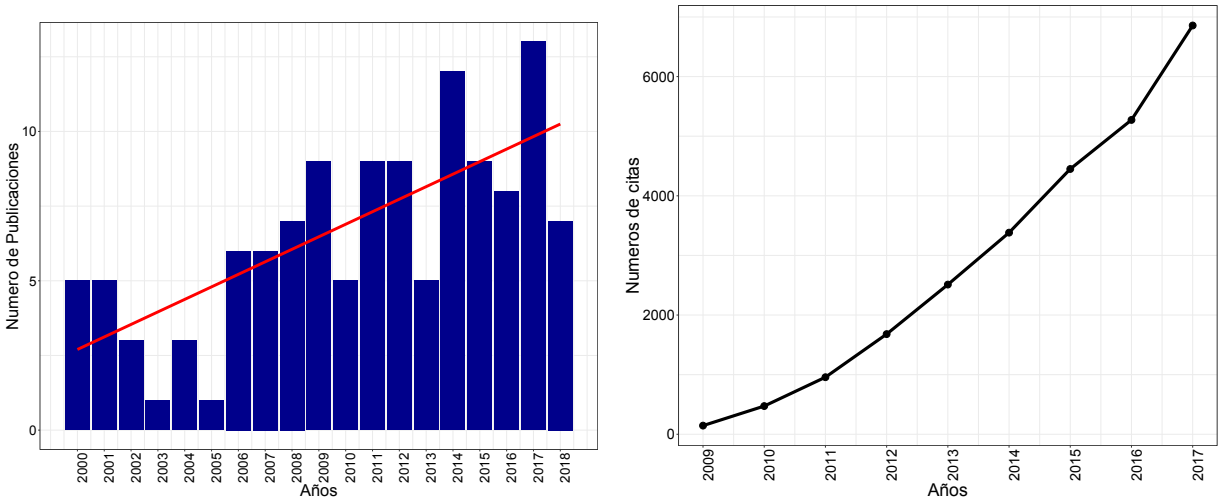
2 Introducción

El uso del término "geocomputación" comenzó en 1996 con la primera Conferencia Internacional sobre "Geocomputación", organizada por la escuela de geografía de la Universidad de Leeds, lanzando una nueva agenda de investigación en análisis geográfico y modelización (Openshaw & Abrahart, 1996). El desarrollo de la computación a gran escala provocado por el avance tecnológico ha contribuido a incrementar la importancia del término geocomputación (Batty, 2017). Una búsqueda rápida del término 'geocomputación" en la base de datos Dimensions [1]demuestra la importancia del crecimiento de este nuevo campo científico, pudiéndose encontrar 4212 publicaciones entre los años 2009 y 2017. En la figura 2.1(a), se muestra la evolución anual de las publicaciones, pudiéndose identificar una tendencia creciente lineal (trazada en rojo) a lo largo del periodo. En términos de impacto, un promedio de 6,1 citas por publicación y un total de 25728 citas (Figura 2.1(b)).

Se han propuesto varias definiciones para el término geocomputación. Rees and Turton, 1998la definen como "*el proceso de aplicación de la tecnología informática a los problemas geográficos*". Para Couclelis, 1998, "*la geocomputación es el universo de las técnicas computacionales aplicables a los problemas espaciales*". Openshaw, 2014 considera la geocomputación como un nuevo paradigma" o "*la aplicación de un paradigma de la ciencia computacional para estudiar todo tipo de fenómenos geofísicos incluyendo los sistemas físicos y humanos*". Cheng et al., 2012 también definen la geocomputación de un modo simple y útil, como "*el arte y la ciencia de la solución de problemas espaciales complejos con ayuda de los ordenadores*". Por lo tanto, la geocomputación no consiste simplemente en aplicar métodos informáticos para explorar conceptos geográficos, sino que ofrece un amplio conjunto de herramientas para el examen y la identificación de nuevas perspectivas sobre los procesos espaciales (Cheng et al., 2012).

A pesar de la ausencia de consenso sobre una definición única de la Geocomputación, es posible identificar cuatro áreas de estudio principales: (1) arquitectura y diseño informático; (2) búsqueda, clasificación, predicción y modelozación; (3) descubrimiento del conocimiento y (4) visualización (Cao et al., 2014; Cheng et al., 2012; Fischer & Leung, 2001; Gahegan, 1999; Openshaw & Abrahart, 1996). En primer lugar el área de arquitectura y diseño informático, aborda las mejoras en el rendimiento informático y el desarrollo de nuevas arquitecturas de

[1]https://www.dimensions.ai/



(a) Numero de publicaciones (b) Numero de citas

FIGURE 2.1: Crecimiento del tema Ceocomputación
Fuente: Elaboración propia

software. En segundo lugar el área de búsqueda, clasificación, predicción y modelización — está relacionada con el progreso en las herramientas de reconocimiento de patrones, clasificación y herramientas de "aproximación de funciones". En tercer lugar, el descubrimiento de conocimiento avanza con los mecanismos mediante los cuales el conocimiento puede destilarse de grandes bases como resultado del uso de herramientas de minería de datos y descubrimiento de patrones. Por último, la visualización aborda los avances en la visualización como un medio de exploración de datos, proporcionando nuevas herramientas y enfoques que proporcionan información sobre base de datos complejos y multidimensionales (Gahegan, 1999). Estos temas no son mutuamente excluyentes; están estrechamente interrelacionados. Se podría incluso argumentar que las áreas de búsqueda, clasificación, predicción y modelización y el área de descubrimiento de conocimiento difieren sólo en los algoritmos utilizados (Cheng et al., 2012).

Adicionalmente, La geocomputación está relacionada con temas como el "big data" espacial (Li et al., 2016; Thakuriah et al., 2017), "machine learning" (Marjanović et al., 2018; Pradhan, 2013; Rogan et al., 2008), computación paralelizada (Guan et al., 2018; Tang et al., 2018), modelos basados en agentes computacionales (Patel et al., 2018; Ramírez et al., 2014; Wang, 2018), desarrollo y aplicación de software (Bivand, 2014; Guan et al., 2018; Rey, 2015) entre otros temas.

Puesto que el enfoque de geocomputación es el análisis espacial (Fischer, 2006), el campo de la economía espacial y especialmente la ciencia regional ha adoptado sus técnicas para estudiar las complejas estructuras de los sistemas urbanos y regionales (Thill & Dragicevic, 2018). Las metodologías tradicionales para la geocomputación, tales como la simulación

basada en agentes y la microsimulación, han sido efectivas en la interconexión de principios computacionales con las teorías fundamentales de la conducta, social y económica para avanzar en la comprensión de las complejas organizaciones de los sistemas regionales (Thill & Dragicevic, 2018).

La tabla 3.1 muestra los resultados de una búsqueda rápida y no intensiva en la Web of Science de tres temas asociados con geocomputación y el campo y área de la economía entre los años 2000 y 2017. Seleccionamos del total de los artículos y capítulos de libros encontrados aquellos asociados a la ciencia urbana y regional y calculamos su participación en área. Los tres temas seleccionados presentan un crecimiento en el número de publicaciones en los últimos 5 años y la participación de la ciencia urbana y regional se encuentra sobre el 9% en todos los años revelando un fuerte vínculo entre la gocomputación y y la economía espacial hasta el punto de proponer un nuevo paradigma fentro de la Ciencias de la información geográfias. (Thill & Dragicevic, 2018; Vaz, 2016, 2018)

TABLE 2.1: Numero de temas de la Geocomputación en el área de la Economía.

Tema	Area	2013	2014	2015	2016	2017	Periodo Total
Modelos Basados en Agentes	Total en el área	68	79	96	116	168	1012
	Urbano/Regional	7	8	17	11	22	128
	Participación	10%	10%	18%	9%	13%	13%
Big Data	Total en el área	26	28	67	73	99	418
	Urbano/Regional	11	12	36	34	38	172
	Participación	42%	43%	54%	47%	38%	41%
Machine Learning	Total en el área	3	3	4	12	36	69
	Urbano/Regional	3	2	1	8	21	37
	Participación	100%	67%	25%	67%	58%	54%

Fuente: Elaboración propia

La tesis doctoral tiene como objetivo contribuir al uso de la geocomputación en el análisis económico espacial mediante la construcción y aplicación de un nuevo conjunto de algoritmos y funciones en el lenguaje de programación R para analizar datos económicos espaciales. La tesis aborda tres temas diferentes en la economía espacial, que requieren el desarrollo de nuevas herramientas geocomputacionales para su análisis(3.2). La tesis hace así una doble contribución, al análisis económico espacial y al paradigma geocomputacional. Contribuye al tema de búsqueda, clasificación, predicción y modelado en geocomputación desarrollando dos paquetes R, el paquete "DataSpa" (Vallone et al., 2017) para recopilar información espacial y el paquete "msp" (Vallone, 2018) para mejorar la calidad de los datos. Contribuye al descubrimiento del conocimiento en el geocómputo mediante el desarrollo de una nueva implementación de herramientas de análisis exploratorio de datos espaciotemporales en el paquete "estdaR" (Vallone et al., 2018). Finalmente, contribuye a la visualización geocomputacional mediante la adaptación de técnicas econométricas espaciales clásicas a la

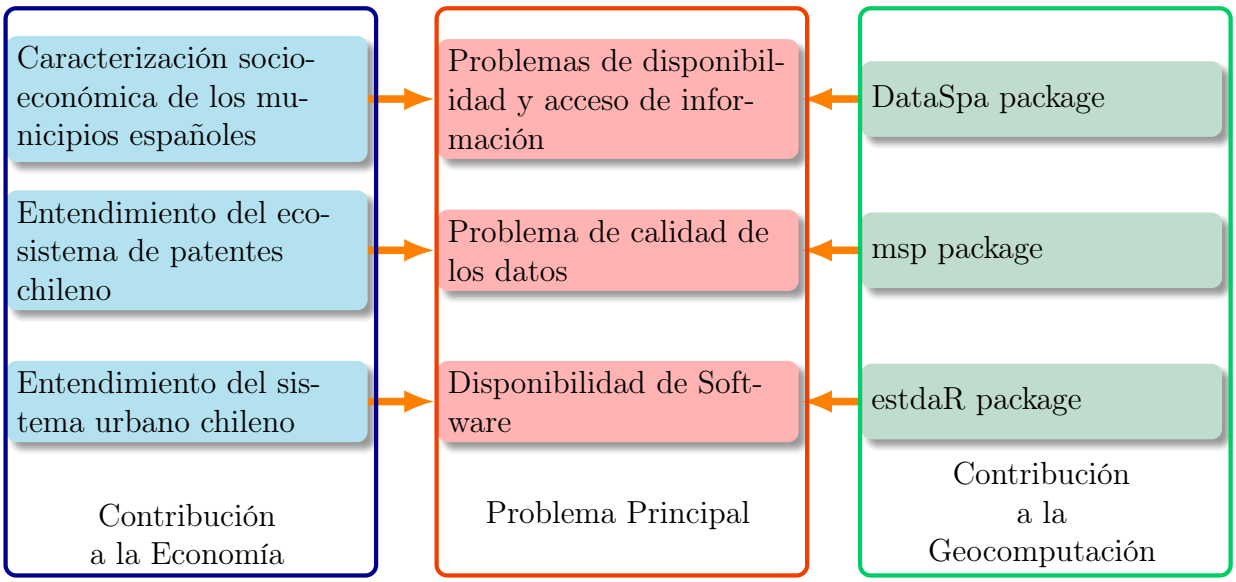


FIGURE 2.2: Contribución de la tesis
Fuente: Elaboración propia

visualización en redes para explorar la existencia de derrames de productividad entre los miembros de una red. La aplicación de estas herramientas a tres casos concretos nos permite contribuir a la literatura de la económica espacial aumentando el conocimiento de la situación socioeconómica de los municipios españoles, incrementando el conocimiento del ecosistema de patentes chileno y aumentando el entendimiento de la dinámica del sistema urbano.

2.1 Temas de investigación

Esta tesis doctoral aborda una serie de temas de investigación que abarcan el proceso típico del análisis económico, desde la recolección de datos, la limpieza y la visualización hasta el análisis estadístico. En primer lugar, considera los problemas de accesibilidad y disponibilidad comúnmente presentes en la recolección de datos basada en la web, dependiendo de la categoría de datos y el país, ya que la información en línea a menudo sólo se proporciona a través de interfaces web parroquiales y difíciles de usar (Foster, 2011). Este es el caso de algunas bases de datos oficiales y no oficiales de información socioeconómica publicadas en España a baja escala espacial, a las que el acceso está restringido o incluso bloqueado. Para superar algunos de estos problemas, el capítulo 3 de la tesis presenta diversas estrategias basadas en el análisis de URL, extracción de texto PDF y raspado web para extraer y organizar varias bases de datos en España a nivel de municipio. Cada estrategia consiste en un conjunto de funciones construidas en el paquete R "DataSpa". En segundo lugar, las

grandes bases de datos suelen presentar problemas de calidad. La exactitud de los datos puede verse afectada por un error humano en la construcción de bases, como errores ortográficos y repetición de nombres. El capítulo 4, una gran base de datos de patentes en Chile, se limpia por medio de un algoritmo de búsqueda y reemplazo "msp", construido en R, que armoniza la información de la base de datos basada en nombres de empresas y evitando errores ortográficos y duplicaciones.

En tercer lugar, una vez que los datos se recopilan, limpian y organizan, puede ser difícil explorar y visualizar bases altamente dimensionales en resoluciones más finas (Brunsdon & Singleton, 2015). Sin embargo, es crucial encontrar patrones, estructuras y relaciones subyacentes interesantes para proporcionar soporte de decisión confiable (Thill & Dragicevic, 2018). La tesis aborda este problema proponiendo nuevas rutinas en R para explorar la existencia de derrames de productividad en la red chilena de co-afiliación de patentes. El capítulo 5 también propone una colección de herramientas para el análisis exploratorio de datos espaciotemporales, aplicada al estudio de la dinámica del sistema urbano chileno durante el siglo pasado.

Por último, la tesis doctoral considera la creciente importancia de la construcción y publicación de paquetes de código abierto en el análisis económico espacial. Los investigadores tienden a desarrollar scripts o funciones para asegurar la reproducibilidad, un problema actualmente importante para los analistas de datos (Brunsdon, 2014). Compartir códigos y funciones relacionadas con temas similares aumenta la visibilidad y la difusión de un estudio específico. Como se explicó anteriormente, un objetivo de esta tesis doctoral es desarrollar herramientas de código abierto para el análisis de bases de datos económicas espaciales. Con este fin, el capítulo 5 presenta un nuevo paquete de R, "estaR", que compila herramientas estadísticas para el análisis espaciotemporal de la evolución de la ciudad en Chile.

2.2 Relevancia de la investigación

2.2.1 Capitulo 3 (Paper 1)

El capítulo 3 tiene como objetivo resolver los problemas relacionados con la recolección de datos a bajos niveles geográficos para estudios urbanos propuestos en España. Los organismos oficiales proporcionan datos sobre las variables de población, fuerza de trabajo y flota de vehículos en España a través de sus portales web–el Instituto Nacional de Estadística (INE)[2], el Servicio Público de Empleo Estatal (SEPE)[3] y la Dirección General de Tráfico

[2]http://www.ine.es.
[3]http://www.sepe.es.

(DGT)[4] respectivamente. Sin embargo, estos conjuntos de datos son difíciles de acceder. La sofisticada estructura organizativa de los portales web de las agencias oficiales, la ausencia de APIs adecuadas y/o la falta de codificación de los datos hacen que el proceso de descarga sea arduo y altamente vulnerable a errores humanos. Para superar estos problemas de accesibilidad, usamos la metodología de URL parsing para analizar las URLs de los portales web de las agencias oficiales y construir un conjunto de funciones de R para conectar con el servidor de la agencia, descargar y manipular la información requerida. Además, dado que la información de DGT está disponible en formato PDF, aplicamos una estrategia de PDF extration para crear un conjunto de funciones de R que cargan y manipulan la información sel parque automotor. En España, los datos de empresas a nivel micro también presentan problemas de disponibilidad. El Directorio Central Empresas del INE (DIRCE)[5] proporciona información anual sobre el número y la distribución de empresas y establecimientos, pero sólo a nivel de meso y macro para preservar la confidencialidad estadística. El Registro Integrado de Establecimientos Industrial (REII))[6], publicado por el Ministerio de energía, turismo y agenda digital, contiene un censo de establecimientos industriales en España a nivel municipal y superior, pero este censo es incompleto para algunas regiones y sectores industriales, sólo incluye empresas que operan en el sector industrial y tiene algunas restricciones de descarga. Para superar estos problemas y crear una empresa y una base de datos de autónomos para España a nivel municipal y superior, utilizamos la técnica de web scraping raspado para construir un conjunto de funciones R para navegar y extraer información en línea proporcionada libremente por la consultora privada Axesor[7]. Todas las funciones forman parte del paquete de R "DataSpa", disponible bajo licencia GPL-2 (Vallone et al., 2017).

2.2.2 Capitulo 4 (Paper 2)

En el capítulo 4 nos enfocamos en el estudio de la actividad de patentamiento y la colaboración en la investigación en Chile entre 1989 y 2013 usando una base de datos de patentes extraído del Instituto Nacional de Propiedad Industrial de Chile (INAPI). El objetivo de este capítulo es proporcionar una visión cuantitativa general del mapa de patentamiento en Chile durante el período de 1989–2013 y utilizando el análisis de redes sociales (Scott, 2013) explorar los patrones de colaboración en la investigación del país. A diferencia de la mayoría de las economías industrializadas, las bases de datos integrales a menudo son incompletas, y los investigadores deben recopilar información directamente en cada caso. En el caso particular de Chile, los datos brutos presentan discrepancias tales como errores

[4] http://www.dgt.es.
[5] http://www.ine.es/dynt3/inebase/es/index.htm?padre=51&dh=1
[6] http://www.minetad.gob.es/industria/RII/Paginas/Index.aspx
[7] https://www.axesor.es.

ortográficos, acrónimos, abreviaturas y nombres registrados de manera diferente. También se encontraron diferentes nombres de empresas que se refieren al mismo individuo o entidad. Algunos nombres de empresas cambian con el tiempo, siendo necesario rastrearlo para modificarlo. Debido a los datos brutos tiene más de 42,000 registro, corregir estos problemas de de manera manual un ineficaz y muy expuesto a fallas humanas, por esa razón basados en un algoritmo de búsqueda y reemplazo construimos una función de R interactiva para resolverlos. Esta función forma parte del paquete "msp" también disponible bajo licencia GPL-2 (Vallone, 2018). Para analizar el ecosistema de patentes chileno y su red de colaboración de investigación, construimos una red de afiliaciones (Scott, 2013) donde las empresas o titulares de patentes son actores y las patentes son los eventos . Además, transformamos esta red de afiliación bimodal en una red unimodal y usamos su matriz de adyacencia como una matriz de "vecindad" en el cálculo de las pruebas de autocorrelación I de Moran y global G para contrastar la relación entre la productividad y colaboración en la red. Para explorar la existencia de un clusters locales productivos en la red, la representamos clasificando a las empresas en cuatro categorias mutuamente excluyentes basadas en los cuatro cuadrantes del diagrama de dispersión de Moran. En general, nuestros resultados muestran la falta de colaboración entre las empresas industriales y la ciencia e industria en Chile, al menos en términos de cotitularidad de patentes. La evidencia revela una intensificación del contenido del conocimiento fuera de las empresas nacionales y extranjeras, pero también una pequeña contribución de las universidades a ese conocimiento. También encontramos que los retrasos en las concesiones son largos y variados. Esta variación es impulsada por las diferencias entre los campos tecnológicos, la experiencia, la residencia y el tipo de empresa. El código de R utilizado en el calculo de este capitulo está disponible gratuitamente para otros investigadores interesados en usarlo.

2.2.3 Capitulo 5 (Paper 3)

El objetivo del capítulo 5 es centrarse en mejorar el conocimiento del sistema urbano chileno, a través de un conjunto de herramientas novedosas que permite evaluar la influencia de la proximidad espacial entre los asentamientos humanos en la evolución de las ciudades para detectar diferencias en estas dinámicas espacio-temporales. Queremos responder a las siguientes preguntas: en primer lugar, ¿existen procesos urbanos homogéneos en todo el sistema de la ciudad chilena? En segundo lugar, ¿es la hipótesis de Duranton (Duranton, 2016) sobre la existencia de efectos de aglomeración más fuertes en los países menos desarrollados aplicables también a Chile? En tercer lugar, ¿en qué medida la población de la ciudad crecerá más rápido o más lento dependiendo de la velocidad de crecimiento de sus vecinos?. Para responder a estas preguntas, usamos los datos del censo durante el período 1930-2002 y primero analizamos la distribución transversal de la población urbana mediante el análisis estadístico

estándar y las estimaciones no paramétricas de las funciones de densidad para algunos años, según lo propuesto por Quah, 1996 y seguido por muchos otros autores (por ejemplo, Xu and Zhu, 2009 y Xiufang et al., 2015). En segundo lugar, el proceso de crecimiento se modela como una cadena de Markov estacionaria de primer orden y el papel del espacio geográfico en las probabilidades de transición se evalúa con un conjunto de métodos basados en una versión espacial de la cadena de Markov estándar (Le Gallo & Chasco, 2008). Tercero, también realizamos un análisis en profundidad para detectar regímenes espaciales en la dirección del movimiento y la movilidad de clasificación de la distribución urbana chilena. Los enfoques de LISA Markov (Rey & Janikas, 2006) y direccional LISA (Rey et al., 2011) capturan la co-evolución de una unidad espacial con sus respectivos vecinos identificando diferentes regímenes espaciales en la clasificación de la movilidad de la distribución urbana. También estudiamos la existencia de diferencias espaciales en el propio patrón de crecimiento con el Indicador Global de la Asociación en la Movilidad (GIMA) (Rey, 2016). Por último, realiza la la descomposición de la ranking (Rey, 2004) que es una medida de cohesión que permite detectar los movimientos de rango sincrónicos entre los regímenes espaciales. Todo el cálculo se realizó utilizando el paquete de R " estdaR" (Vallone et al., 2018) disponible bajo licencia GPL-2. Los resultados muestran la existencia de diferentes dinámicas regionales, reflejando la existencia de heterogeneidad espacial en el sistema urbano chileno, también hay evidencias de un patrón claro y persistente de las economías de aglomeración en el sistema urbano chileno. Finalmente, la probabilidad de que una ciudad crezca aumenta con el tamaño de sus vecinos, mientras que las grandes ciudades rodeadas de ciudades más pequeñas no experimentarán prácticamente ningún cambio en la población, por lo tanto, las cuestiones de proximidad espacial en el sistema urbano chileno.

3 Some strategies to access web-based urban spatial data for socioeconomic research using R functions

3.1 Introduction

In recent years, there have been an impressive increase in web-based research (Denissen, Neumann, & Zalk, 2010). Nowadays it is common the use of internet-based databases which are obtained by either primary data online surveys or secondary official and non-official registers (Chang, Kayed, Girgis, & Shaalan, 2006; Howard et al., 2015; Siewert & Udani, 2016; Wright, 2005). The Internet has also transformed the way researchers interact with secondary data, reducing the cost of collecting, updating and storing datasets from government agencies. It has also increased the availability of non-structured information in non-official web pages to research (Edelman, 2012; Hooley, Wellens, & Marriott, 2011). However, information disposal varies depending on data category and country (Graham, Hogan, Straumann, & Medhat, 2014). On many occasions, the collection of microdata at low geographical level for urban analysis could become a challenge. In effect, certain data collected for statistical purposes by government agencies from households, individuals and business establishments through census and surveys are never made available due to a pledge of confidentiality restrictions (National Research Council, 2005). Instead, data is provided either in the form of restricted-access data files or as anonymized data products, in which geocoded information is only available at a regional –aggregated– spatial level.

The most common difficulties when working with secondary Internet-based data can be grouped into two categories: accessibility and availability problems. Accessibility problems are present when the way that data is published in the servers, blocks or delays the download process. Then, data collection becomes a tedious reiterative task that can produce errors in the construction process of big databases. Availability problems usually arise when the official agencies restrict access to the information for statistical confidentiality reasons or when data is simply non-existent.

Two elements contribute to reduce these problems. First, the changes produced in the research paradigm due to the increasing use of open-source software, like Python or R, which allows incorporating data collection, manipulation and publication processes into a single software environment (Simon, Christian, Peter, & Dominic, 2015). Second, the increasing use of API technologies and new data collection techniques as the web scraping (Glez-Peña, Lourenço, López-Fernández, Reboiro-Jato, & Fdez-Riverola, 2014; Grasso, Furche, & Schallhart, 2013; Mehlführer, 2009; Nolan & Temple Lang, 2014; Penman, Baldwin, & Martinez, 2009; Salamone, Scannapieco, & Scarnò, 2014; Simon et al., 2015).

In order to overcome some of these problems, this paper presents different strategies based on URL parsing, PDF text extraction and web scraping. These approaches have been used to extract and organize several databases on population, unemployment, vehicle fleet and firm in Spain at the municipality level (LAU)[1], for which accessibility to information is limited and problematic. Each strategy consists of a set of functions built in the R package, "DataSpa"[2], which is available under a GPL-2 license (Vallone et al., 2017). They allow collecting higher quality information by avoiding potential human errors due to different impediments and restrictions imposed by official and non-official web portals to microdata extraction. This package, which constitutes the main contribution of this paper, was built to elaborate the 2017 Socioeconomic Atlas of Extremadura, which constitutes the most important official database of municipality variables in this region. "DataSpa" is very useful for not only the researchers interested in the analysis of urban systems in Spain but also, by a convenient adjustment of the package functions, any other analysts who face similar problems in other countries.

This chapter is organized as follows: after the introduction, section 3.2 presents URL parsing as a suitable strategy to download and encode population and unemployment databases, for which a sophisticated publication platform creates serious accessibility problems. In section 3.3, we illustrate the performance of the PDF extraction strategy with the case of the vehicle fleet database, in which the absence of an API and some blocking systems lead us to download the PDF files with the municipality reports to extract the available information. Section 3.4 presents the use of web scraping to download a database of firms published by a private company, which helps to solve the absence of this information in Spain at the urban and individual level. In section 3.5 we show the use of the package in the 2017 Socioeconomic Atlas of Extremadura and in section 3.6 we present the conclusions.

[1] According Eurostat, The LAUs (Local Administrative Units) are subdivisions of the NUTS 3 regions, which consist of municipalities or equivalent units (formerly NUTS 5). The NUTS classification (Nomenclature of territorial units for statistics) is a hierarchical system for dividing the economic territory of the EU. NUTS 1 are major socio-economic regions (e.g. Spain), NUTS 2 are basic regions for the application of regional policies (e.g. autonomous community of Extremadura) and NUTS 3 are small regions for specific diagnoses (e.g. province of Badajoz).

[2] This R package is freely available from the site https://github.com/amvallone/DataSpa. It must be installed in the R console with the command: `devtools::install_github("amvallone/DataSpa")`.

3.2 URL parsing for databases with accessibility problems

In many cases, information is available in statistical sites, but the sophistication of their organizational structure, the absence of adequate APIs and/or the lack of codification of the data make the downloading process tiresome and very exposed to human failures. This is the case of two main variables for socioeconomic analysis like population or workforce in Spain, which are provided by two official agencies in their web portals, the National Statistics Office (INE)[3] and the Spanish Public Employment Service (SEPE)[4]., respectively, but with certain access barriers. By way of illustration, downloading of data population for the municipalities of only one of the 52 Spanish provinces requires clicking at least five different URL links to reach the final data in a webpage. The construction of a whole panel data on population, which is available on the INE website since 1996, for more than eight thousand municipalities and different categories (population by sex, age group and nationality) is a long-term task. In the case of unemployment, the SEPE web server contains municipality data on registered working contracts and unemployed population. The downloading conditions for these variables are even worse than in the previous case: to the difficulty of clicking at least five times to reach the final database for each province, we have to add that municipality data is only provided as monthly series and, before 2012, without the official municipality identification code. Hence, the operating time for data extraction is even higher in the SEPE web portal, which presents the extra obstacle of having to code the municipalities for cross-matching tasks. It must be said that either INE or SEPE admit custom demands of large volumes of data but they are not always satisfied or free, depending on the requesting institution, and they usually involve certain amount of discouraging paperwork.

In order to overcome these problems, we have created a code to download and manipulate the population and workforce municipality data, which are freely available in the respective official agencies but rather limited, as shown previously. The construction strategy is based on the "Uniform Resource Locator" (URL) parsing method, which consists of taking a URL in order to break it out in its standard components: scheme, domain, port, path, query string and fragment identifier to extract information about the abstract or physical resource associated with it (Berners-Lee & Masinter, 2015). This method, which belongs to the computational science, has been used to avoid financial phishing attacks (Liu & Zhang, 2012; Zuhair, Selamat, & Salleh, 2016) , evaluate security problems (Bhargavan, Delignat-Lavaud, & Maffeis, 2013), analyze the user's cost-benefits of ignoring the security adviser (Herley, 2009) and retrieve images using Web mining (Chen, Wenyin, Zhang, Li, & Zhang,

[3]http://www.ine.es.
[4]http://www.sepe.es

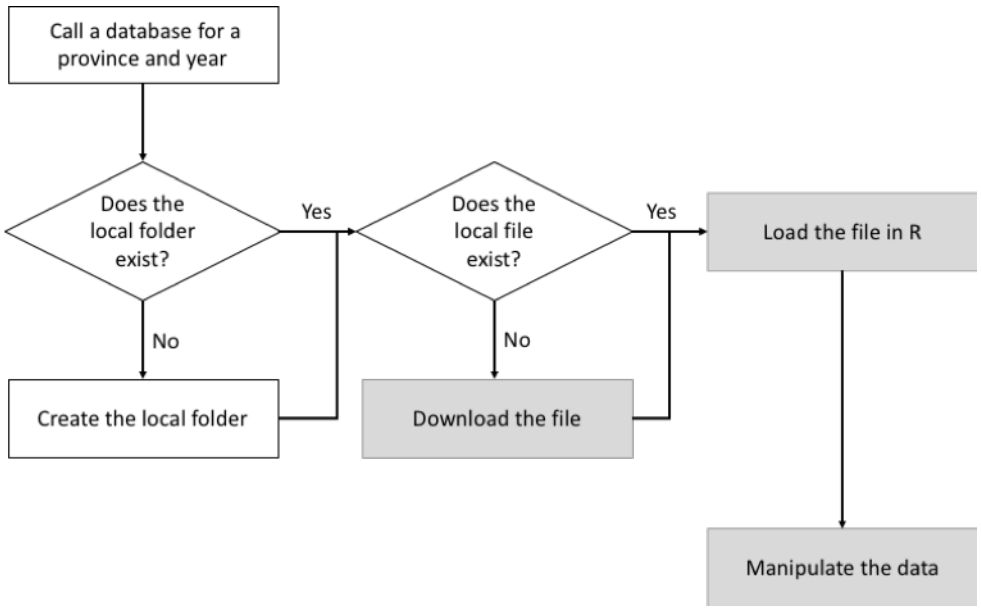


FIGURE 3.1: Workflow of the URL parsing functions to download databases with accessibility problems
Source: self-elaboration

2001) , among others. In a similar process than the used by this later paper, we apply text analysis over the URL path. The path of a URL is made of text segments that represent a structured hierarchy, similar to a directory structure, where each segment is separated by the "/" character. The common path segments can be detected from the analysis of the text to identify different category segments corresponding to characteristics of the database (sex, year, province, etc.) With this information, it is possible reproduce the specific URL of the file containing a desired dataset, which will be available to subsequent downloading and manipulation.

Figure 3.1 presents the general workflow this URL parsing functionality designed for the databases on population and unemployment in Spain. Based on the province and year of the user's interest, a set of R functions are implemented. At the beginning of the execution, the code checks the existence of a folder to host the download file. If this element does not exist, a folder named `"data_poblacion"` or `"data_paro"` is created depending on the required data. Then, it looks for the existence of the file in the local host. If this file does not exist there, it will be downloaded from the server and hosted with a determined name, depending on the function. The creation of a local data store facilitates its accessibility because it reduce Internet and URL dependencies. In effect, a local data store allows using the database without an Internet connection and prevents from changes in URLs. Additionally, it improves the code performance in the last part of the process, reducing the user's elapsed time to import the data into R.

We have built a set of nested functions to increase the code flexibility. Each grey box in Figure 3.1 corresponds to one of the following function subsets: download functions, loading functions and data manipulation functions.

3.2.1 URL parsing download functions

The main purpose of the download functions is connecting to the agency server to download a requested file. These functions have a simple two-step process. First, depending on some given arguments, they check the existence of a local folder, which will be created –when not available– to store the file. Second, they create the URL address connected to the file and download the file. Next, we present a brief description of the download functions used to extract the municipality data on population and workforce from the INE and SEPE web portals:

(1) `getbase.pob(year,"provincia",extr=FALSE,anual=FALSE)`: This function downloads a file with the municipality population data by sex and five-year age for a Spanish province referred to a determined year. It calls the file `"pob_q_year_provincia.xls"` and saves it in the folder `"data_poblacion"`. This function has four arguments. `year` is a numerical value for the reference year of the dataset. `"provincia"` corresponds to each of the 52 Spanish provinces. `extr` ("*is foreign population?*") is a logical variable with `FALSE` as the default value, for which `extr=TRUE` downloads and saves the foreign population dataset by sex and major group-year age. `anual` ("*is data required by age?*") is a logical variable with `FALSE` as the default value, for which `anual=TRUE` downloads and saves the population by sex and one-year age. Since there is not data for foreign population by one or five-year age, the combination of `extr=TRUE` and `anual=TRUE` will generate an error message ("No data for these cases"). For example, the command `getbase.pob(2016,"Badajoz")` downloads municipality population data of the province of Badajoz corresponding to the year 2016 in the file `"pob_q_2016_BADAJOZ.xlsx"`.

(2) `getbase.fen(year,"provincia")`: This function downloads a file with other municipality demographic data (live births, fetal deaths, marriages, etc.) for a Spanish province referred to a determined year. It calls the file `"fen\_year\_provincia.xls"` and saves it in the folder `"data\_poblacion"`. This function has two arguments: `year`, which is a numerical value for the reference year of the dataset and `"provincia"`, which corresponds to each of the 52 Spanish provinces. For example, the command `getbase.fen(2016,"Badajoz")` downloads municipality demographic data of the Badajoz corresponding to the year 2016 in the file `"fen_2016_BADAJOZ.xlsx"`.

(3) `getbase.paro(year,"mes","provincia")`: This function downloads a file with the municipality unemployment data by sex of a Spanish province referred to a period of time. It calls the file `"paro_MUNI_provincia_mmyy.xls"` and saves it in the folder `"data\_paro"`. The function has three arguments: `year`, which is a numerical value for the reference year of the dataset, `"mes"`, which is the value for the reference month of the dataset and `"provincia"`, which corresponds to each of the 52 Spanish provinces. For example, the command `getbase.paro(2016,"julio", "Badajoz")` downloads municipality unemployment data of Badajoz corresponding to the month of July of the year 2016 in the file `"paro_MUNI_BADAJOZ_0716.xls"`. As an example of URL parsing download function, we illustrate this routine in Algorithm 3.1.[5]

ALGORITHM 3.1: URL parsing download function *getbase.paro()*

```
# Example of download function used to extract the Spanish
   municipality data on workforce
# Output: The Ms. Excel file  paro _MUNI_provincia_mmyy.xls
   , which is saved in the folder  data _paro .

getbase.paro<-function(year,mes,provincia){
  year<-as.character(year)
  if(dir.exists(file.path(getwd(),"data_paro"))==FALSE){
    dir.create(file.path(getwd(),"data_paro"))
    }
  provincia<-toupper(provincia)
  provincia<-a.letter(provincia)
  mes<-tolower(mes)
  nn.mes<-seq(1,12,1)
  names(nn.mes)<-c("enero","febrero","marzo","abril","mayo","
   junio","julio","agosto","septiembre","octubre","noviembre","
   diciembre")
  cod<-paste("0",nn.mes[mes],substr(year,3,4),sep="")
  name<-paste(paste("MUNI",provincia,cod,sep="_"),".xls",sep=""
   )
  url<-paste("http://www.sepe.es/contenidos/que_es_el_sepe/
   estadisticas/datos_estadisticos/municipios/",year,"/",paste(
   mes,year,sep="_"),"/",name,sep="")
  dir<-paste(getwd(),"/data_paro/",sep="")
```

[5] All the R functions are in the aforementioned repository: *https://github.com/amvallone/DataSpa*.

```
  file<-paste(dir,"paro_",name,sep="")
  download.file(url,file, mode='wb')
}
```

3.2.2 URL parsing loading functions

The mean purpose of the loading functions is importing into R the already downloaded and stored databases. These functions have a simple two-step process. First, depending on the given arguments, they check the existence of a required file in the local folder. Second, if the file does not exit, they call the corresponding download function to create it, using the 'xlsx' R package to import the file. These functions are the following:

(1) `paro(year,"mes","provincia")`: This function has the same arguments than the already shown `getbase.paro()` function. The output of this function is a data frame containing the following variables: official municipality identification code ("cod"), municipality name ("nombre"), number of unemployed people in the municipality ("paro total"), number of unemployed males ("hombres") and number of unemployed females ("mujeres"). For example, the command `paro(2016,"julio","Badajoz")` generates a data frame containing the unemployment database corresponding to the municipalities of Badajoz in July 2016.

(2) A set of nine functions, which share the same arguments, but produce different outputs. Each function has two arguments: `year`, which is a numerical value indicating the year of the requested data and `"provincia"`, which is one of the 52 Spanish provinces. Notwithstanding the functions have a different output, there are two common variables listed by default: the official municipality identification code and the municipality name. Next, we present a brief description of these outputs by function. Functions `pob.a()` and `pob.q()` produce three data frames, all of them containing total population by sex and age (one-year and five-year age groups, respectively). `pob.e()` creates a list of three data frames, all of them containing total population and population by age (major groups) and nationality (nationals and foreigners), for both sexes, males and females. `pob.tot()`, `pob.h.tot()` and `pob.m.tot()` generate, each one, a data frame containing municipality data for total, male and female population, respectively. `pob.n.tot()` and `pob.e.tot()` generate, each one, a data frame containing municipality data for total national and foreign population, respectively. `pob.fen()` generates a data frame containing municipality data for the number of births and deaths. As an example of URL parsing loading function, we illustrate this last routine in Algorithm 3.2.

ALGORITHM 3.2: URL parsing loading function *pob.fen()*

```
# Example of loading function used to generate a data frame of
   Spanish municipality data on births and deaths
# Input: a URL parsing download function getbase.fen(year,
   provincia) of the DataSpa package.
# Output: an R data frame.
pob.fen<-function(year,provincia){
    year<-as.character(year)
    provincia<-toupper(provincia)
    prov<-provincia
    provincia<-a.letter(provincia)
    dirc<-paste(getwd(),"/data_poblacion/",sep="")
    file<-paste(paste("fen",year,provincia,sep="_"),".xlsx",sep
   ="")
    if(sum(dir(dirc)==file)==0){
      getbase.fen(year,provincia)
    }
    abre<-paste(dirc,file,sep="")
    datos<-xlsx::read.xlsx(abre,1,colIndex=c(1,2,5))
    datos<-datos[which(complete.cases(datos)==TRUE),]
    datos<-datos[-1,]
    d<-dim(datos)
    nombres<-as.character(datos[,1])
    codigo<-rep("AA",d[1])
    municipio<-rep("AA",d[1])
    for (i in 1:d[1]){
      nn<-unlist(strsplit(nombres[i]," "))
      codigo[i]<-str_trim(nn[1])
      if(length(nn)>2){
        nom<-paste0(nn[2:length(nn)],collapse=" ")
        municipio[i]<-nom
      } else {
      municipio[i]<-str_trim(nn[2])
      }
    }
    cifras<-as.data.frame(datos[,2:3])
    cifras<-apply(cifras,2,as.numeric.factor)
    ids<-as.data.frame(cbind(codigo,municipio))
```

```
    salida<-cbind(ids,cifras)
    colnames(salida)<-c("Cod","Municipio","Nacidos","Fallecidos
  ")
    salida
}
```

3.2.3 URL parsing manipulation functions

These functions have the purpose of manipulating the data to build space-time panels of municipality variables for different periods or compute demographic indicators. These new variables will be stored as R data frames and/or Ms. Excel output files. All functions employs either a download or a loading function. Next, we present the manipulation functions included in the "DataSpa" package.

(1) A set of six `.ev()` functions for the construction of panels of population variables at the municipality level for a time period.[6] All these functions share the same three arguments: `inicio`, which is the starting year of the panel, `fin`, which is the last year of the panel and `"provincia"`, which is one of the 52 Spanish provinces. Although these functions also have a different output, there are four common variables listed by default: the official municipality identification code, the municipality name and the columns corresponding to the initial and final years of the population panel. `pob.ev()`, `pob.h.ev()` and `pob.m.ev()`, `pob.n.ev()`, `pob.e.ev()` generate, each one, a data frame containing a panel of municipality variables for total population, males, females, nationals and foreigners, respectively, for a given time period. `pob.fen.ev()` produces two data frames, all of them containing the output elements (municipality codes, names and panel variables), for births and deaths. For example, `pob.ev(2000,2016,"Badajoz")` generates a data frame containing the total population corresponding to the municipalities of Badajoz for the period 2000-2016. As an example URL parsing manipulating function, we illustrate the `pob.ev()` routine in Algorithm 3.3.

(2) `pob.ind(year,"provincia",print=FALSE)` compute a data frame with a set of demographic indexes at the municipality level. It has the following arguments: `year`, which is a numerical value indicating the year of the requested data, `"provincia"`, which is one of the 52 Spanish provinces and `print`, which is a logical variable with

[6] These functions deal with two important difficulties derived from the construction of panels for municipality data in Spain. First, they control for municipality entries and removals, which take place almost every year, adapting the final data frame to the configuration corresponding to the last period. Second, they produce a list of name equivalences, based on the information provided by the INE, to manage with constant changes in the municipality names, always assigning the one corresponding the last period.

FALSE as the default value, for which `print=TRUE` saves the dataset as a Ms. Excel file. The output of this function is a data frame containing a set of ten demographic indexes: Childhood, youth, third age, dependence, unemployment rates (both sexes, males and females) and municipality average age (both sexes, males, females). There is also a similar function `pob.ind.p()`, which computes finer indexes using one-year age groups (instead of five) from 2011.

(3) `ind.ev(inicio, fin,"provincia",print=FALSE)`. It has four arguments: `inicio`, which is the starting year of the panel, `fin`, which is the last year of the panel, `"provincia"`, which is one of the 52 Spanish provinces and `print`, which is a logical variable, with `FALSE` as the default value, for which `print=TRUE` saves the dataset as a Ms. Excel file called `"pob_ev_index_provincia_inicio-fin.xlsx"`. The output creates ten data frames, all of them containing the municipality code and name and the requested time series, for each of the demographic indexes obtained with the `pob.ind()` function.

ALGORITHM 3.3: URL parsing manipulation function *pob.ev()*

```
# Example of manipulation function used to generate a data
   frame of Spanish municipality data on population for a given
    time period
# Input: a URL parsing download function  getbase.fen(year,
   provincia) of the    DataSpa    package.
# Output: an R data frame.
pob.ev<-function(inicio,fin,provincia,print=FALSE){
  if(fin<inicio) stop("La fecha de inicio debe ser mayor que la
    fecha de fin")
  n<-seq(inicio,fin,1)
  year<-as.character(sort(n,decreasing=TRUE))
  base<-pob.tot(year[1],provincia)
  for (i in 2:length(year)){
    aux<-rep(NA,dim(base)[1])
    pob<-pob.tot(year[i],provincia)
    v<-intersect(base[,1],pob[,1])
    for(j in 1:length(v)){
      aux[which(base[,1]==v[j])]<-pob[which(pob[,1]==v[j]),3]
    }
    base<-cbind(base,aux)
    }
  colnames(base)<-c("Cod","Municipio",year)
```

```
  orden<-c(1,2,seq(dim(base)[2],3))
  base<-base[,orden]
    if (print==TRUE){
      if(dir.exists(file.path(getwd(),"Outputs"))==FALSE){
      dir.create(file.path(getwd(),"Outputs"))
      }
    file<-paste(getwd(),"/Outputs/pob_total_ev_",provincia,"_",
   paste(inicio,fin,sep="-"),".xlsx",sep="")
    xlsx::write.xlsx(base,file)
    }
  base
}
```

3.3 PDF text extracting for databases with accessibility problems

The Portable Document Format (PDF) is a widely used digital document file format. It is designed to allow users to view, print and exchange electronic documents preserving their look, across platforms with different operating systems and hardware environments (Marinai, 2009). Though there are many tools for generating PDF files from text documents, there is no standard tool for converting PDF files into texts with 100% accuracy (Thaiprayoon & Haruechaiyasak, 2016). That is, the PDF format does not easily allow extracting the file information (text, tables, images, databases etc.) in a straightforward way because it is hard to handle (Castillo-Fernández, 2015). A PDF file describes the appearance of a page but do not mark up the logical content. Any transformation of the content of a PDF file into text format will imply a reconstruction of words and sentences from the raw positions of the letters included in the PDF.

Statistical web portals and digital platforms usually offer their databases in different formats (HTML, spreadsheets, etc.) in order to reduce data accessibility problems, such the absence of an API, sever instabilities or limitations in the downloadable records. The problem arises when data is only accessible from non-editable formats, like PDF files. This is –partially– the case of the municipality database on vehicle fleet in Spain. Vehicle fleet is a relevant variable in urban studies, which is used to study, for example, residential location (Eluru, Bhat, Pendyala, & Konduri, 2010) , urban air pollution (Kahn & Schwartz, 2008; Mage et al., 1996; Wang, Fu, Lin, Zhou, & Chen, 2009) and effect on urban structure and commutation choice (Bento, Cropper, Mobarak, & Vinha, 2005). The Spanish National Department of

Traffic (DGT)[7] collects and distributes information about vehicle fleet at the municipality level, but before 2014, there is not an API to access to his server. One single download of municipality information always exceeds the maximum allowed data volume. In addition, a CAPTCHA field must be filled to avoid robot access what adds more time to the data collection process. Hence, the only way of downloading the whole vehicle fleet database at once is extracting the information provided by the DGT in PDF format.

Despite the aforementioned difficulties, there are a set of tools which transform and extract information from a PDF file into readable format (Hadjar, Rigamonti, Lalanne, & Ingold, 2004) . Generally, a PDF data extraction process involves at least two steps: first, it transforms the PDF to a readable file and second, it extracts the demanded information. These tools have been used by administrative services to extract automatically documents in digital libraries (Marinai, 2009) or to extract metadata from scientific articles (Aumueller, 2009; Beel, Langer, Genzmehr, & Müller, 2013). They have also been used in more sophisticated contexts to generate text input for text-mining software in-situ in the Mouse Genome Informatics (MGI) system (Dowell, McAndrews-Hill, Hill, Drabkin, & Blake, 2009) or to extract and classify vectorized diagrams (Futrelle, Shao, Cieslik, & Grimes, 2003).

For the municipality database on vehicles, we have created an R function, which downloads the municipality report in PDF format to extract all the available information on it. We combine URL parsing and PDF extraction methods to create the function. URL parsing is useful to detect and construct the URL that leads to the PDF information. To this aim, we use the "rvest" R package (Wickham, 2016). We construct a web crawler to extract a list of URLs with the download addresses corresponding to each municipality PDF file. The DGT web portal offers its databases in a treelike structure: each NUTS2 region ("comunidad autónoma") has as many URLs as NUTS 3 provinces, annual periods and NUTS5 municipalities.[8] Second, a PDF extraction method is used to extract the information from the PDF file. In this case, we use the free software "PDF2TXT"[9] to transform the PDF file into a editable text format to reconstruct the database using text mining in R.

Figure reff2-2 summarizes the function `parque.aut()` workflow. To reduce the Internet dependencies and the elapsed time, the function creates a local folder to host the files. The first step of the process consists in checking the existence of the local folder: if it does not exist, the function creates it. Each province and year has its own folder into the local host and the files are stored as `"year_name.pdf"`, where "year" is the requested annual period and "name" is the municipality name. In a second step, this function looks for the existence – in the local host– of the municipality reports in text format: if they do not exist, the function

[7] http://www.dgt.es
[8] http://www.dgt.es/es/seguridad-vial/estadisticas-e-indicadores/informacion-municipal
[9] https://www.pdf2txt.com

creates the list of URLs to the PDF reports to download the files into a specific local folder. These PDF files are transformed into plain text and they are stored in text files. Since the PDF files are no longer necessary, they are deleted. When the text files are available, the function imports the files into R and creates a data frame with the municipality information for the requested province.

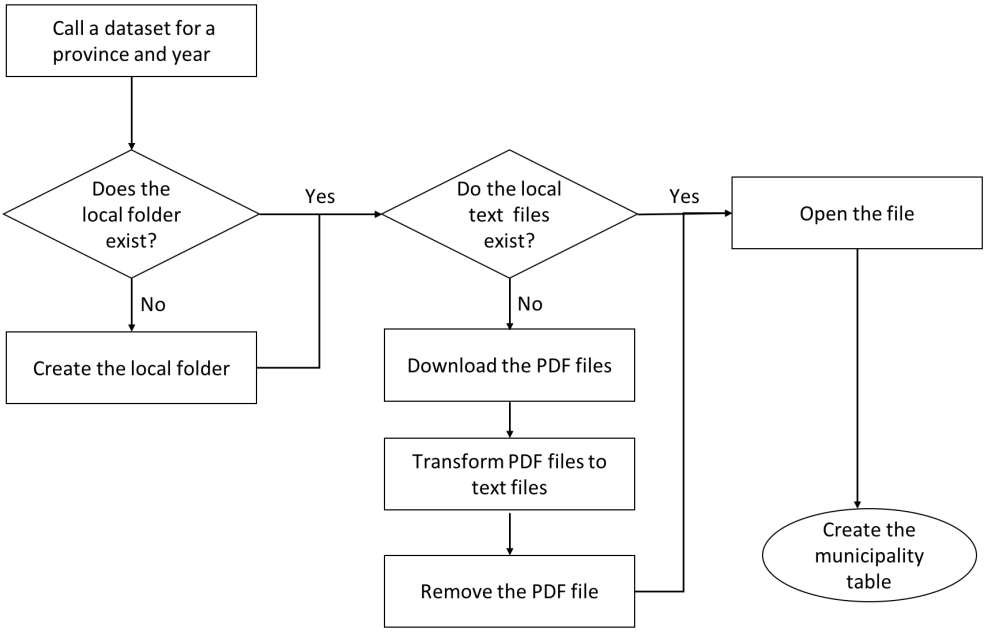


FIGURE 3.2: The function parque.aut() workflow.
Source: self-elaboration

The function `parque.aut(year,"ca","provincia")` has three arguments: `year`, which is numerical variable indicating the year of the requested data, `"ca"`, which is one of the 17 NUTS2 regions ("autonomous communities") and `"provincia"`, which is one of the 52 Spanish provinces. The output of this function is a data frame containing the following variables: the municipality name, number and average age of vehicle fleet, type of vehicles (cars, vans, trucks, motorcycles, buses, etc) and some other variables related to the register of drivers, accidents and vehicle taxes.

3.4 Web scraping for databases with availability problems

Availability problems arise when certain information is not available at all or it is only obtainable partially, which is quite common in micro-territorial databases due to the confidentiality, privacy or any other reasons. In these occasions, the use of alternative information

sources –when existing– is a convenient alternative. In Spain, this is the case of firm data, which is not publicly available at the micro-level. Firm data is a key information in urban studies (location decision, agglomeration economies, entrepreneurship), particularly at firm level or lower spatial aggregations, like census tracts, districts or municipalities (Arauzo Carod, 2005; Arauzo-Carod & Viladecans-Marsal, 2009; Jofre-Monseny, Marín-López, & Viladecans-Marsal, 2011).

The INE's Central Company Directory (DIRCE)[10] provides annual information on the number and distribution of companies and establishments, but only at meso (regional NUTS 2 and 3) and macro-level (national NUTS 1), in order to preserve statistical confidentiality. The Industrial Establishment Register (REII)[11], published by the Ministry of Energy, Tourism and Digital Agenda, contains a census of industrial establishments located in Spain at the municipality level (NUTS 5) and above, but it is not complete for some regions and industrial sectors. It only comprehends companies operating in the industrial sector and it has also some downloading restrictions. Hence, the only way of obtaining firm data for all the economic sectors at the individual level or, at least, aggregations for census tracts, districts or municipalities, is buying them to specialized companies, such as Camerdata[12], the Iberian Balance Sheets Analysis System (SABI)[13], the Global Entrepreneurship Monitor (GEM)[14] or Axesor[15]. All in all, it is not always possible to download the complete database once and for all. For example, SABI has a restricted access to only a set of 50.000 weekly records of Spanish firms.

One of these private consulting firms, Axesor, offers part of its huge database on firms and freelances freely online. Hence, by a web mining process of this information, it is possible to create a database for Spain at the firm-level and above. Web scraping is a software technique which extracts information from websites, usually simulating human exploration of the World Wide Web (Kumar, 2015). Human behavior can be simulated by a web crawler, which is a bot that systematically browses the World Wide Web. It starts with a seed list of URLs to visit in order to identify all the hyperlinks in these pages to add them to a new list of URLs called the crawl frontier. Then, the URLs of the frontier are recursively visited according to a set of policies (Kumar, 2015). Web scraping is focused to transform unstructured or semi-structured data on the web, typically in HTML format, into structured information (Kumar, 2015; Mehlführer, 2009).Typically, this extraction is made by a text mining process, searching for key words and extracting the information associated to them. This technique has multiple uses in economic research (Edelman, 2012); for example, to compute Consumer

[10] http://www.ine.es/dynt3/inebase/es/index.htm?padre=51&dh=1
[11]
[12] http://www.camerdata.es/index.php
[13] https://www.bvdinfo.com/en-gb/our-products/data/national/sabi
[14] http://www.gem-spain.com
[15] https://www.axesor.es

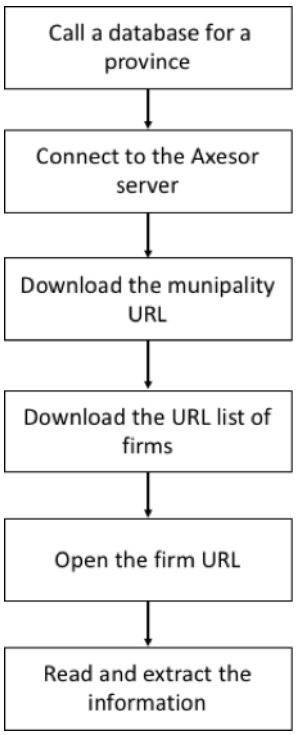


FIGURE 3.3: The function data.firm() workflow.
Source: self-elaboration

Price indexes (Griffioen, de Haan, & Willenborg, 2014; Nygaard, 2015; Polidoro, Giannini, Conte, Mosca, & Rossetti, 2015) and enterprise innovation activities (Gök, Waterworth, & Shapira, 2015). The Axesor database is grouped by province, having each of them its own website with a municipality list. Every municipality, in turn, has a website with a list of firms, which likewise leads to an own website containing extensive information about location, corporate information and several business and financial indicators, for different annual periods. In January 2017, there were about 3,500,000 firms in Spain according to Axesor.

Figure 3.3 shows the workflow of the function `data.firm()` for firms (and `data.firm.a()` for the freelancers). First, for a given province, the function connects to the Axesor server, downloads the municipality URL links, and enters in each of them to download their corresponding firm URLs. Second, every firm URL is opened so as the functions can read and extract the data and construct a data frame. We have followed a design strategy based on two R packages. First, after exploring the Axesor web page HTML code, using the "rvest" package (Wickham 2016) we build a web crawler to obtain the firm URLs. Second, once the firm URLs were accessible, the "stringr" package (Wickham 2017) builds a function to analyze and extract the text information from the web page.

The use of the `data.firm()` and `data.firm.a()` functions is simple because they

only depends on one argument: `data.firm("provincia")` and `data.firm.a("provincia")`, where `"provincia"` is a character variable indicating one of the 52 Spanish provinces. The main difference between both functions is the output contains: while `data.firm()` bring a data frame containing 21 variables, `data.firm.a()` creates a data frame of 12 variables, because the freelanders or self-employed dataset contains less information. The data frame output of the `data.firm()` function contains the following variables for each company: location (province, municipality, address, geographic coordinates), company characteristics (name, birth, legal form, social object), main figures (number of employees, social capital, sales), economic activity codes and firm URL. The `data.firm.a()` function output data frame does not contains the variables of geographical coordinates and main figures.

Since the data collection process is time-consuming, it is possible to divide the whole procedure into a set of functions, allowing an advanced R user to parallelize the process, though it is not recommendable to avoid server crashes. These functions are available and documented in the "DataSpa" package.

Axesor constitutes an interesting alternative to the well-known SABI dataset, though it should be said that some of these variables (mainly the main figures) present incomplete information.

3.5 Case example: the 2017 Socioeconomic Atlas of Extremadura

The "DataSpa" package and its routines was built in order to prepare the 2017 Socioeconomic Atlas of Extremadura (Junta de Extremadura, 2017). This online publication constitutes an extensive compendium of valuable statistical information referred to the region or autonomous community of Extremadura (Spain), which is presented in tables and thematic maps for different spatial scales. Most of the tables are maps contains data of each of the 388 municipalities (LAUs) of Extremadura. There are some indicators and maps for the districts of the seven main municipalities (those with more than 25,000 inhabitants): Almendralejo, Badajoz, Cáceres, Don Benito, Mérida, Plasencia and Villanueva de la Serena. Some tables also present aggregated data for the two NUTS 3 or provinces of Extremadura and 28 commonwealths of municipalities or "mancomunidades" (formerly NUTS 4 or LAU 1).

The Atlas comprehends nine chapters and four annexes including, among others, many indicators of economic activity, demographic phenomena, entrepreneurship and social welfare.

Many of these variables have been downloaded and treated using "DataSpa", as shown in Table 3.1. URL parsing download, loading and manipulating functions have been crucial to generate tables of socio-demographic indicators. This is the case of the following variables:

a) Total population.

b) Population by sex: males and females.

c) Population by age groups: childhood index, youth index, old-age indexes, average age of the population.

d) Population by nationality: nationals and foreigners.

e) Natural population movement: birth, death, fertility and maternity rates.

f) Unemployment: number of unemployed people and unemployment rates by sex.

g) Panels of time-series for many of the previous municipality databases from 2000 to 2016.

A PDF extraction function was employed to download and build tables for vehicle fleet by type (automobiles, motorcycles, vans, etc.) and their corresponding time-series panels from 2000-2015. Finally, two web scrapping functions allowed us to download and build the following variables for entrepreneurship in companies and self-employed: number of entrepreneurial activities, entrepreneurial activity index, sectoral participation rates and local sectorial specialization rates. All these variables are offered by activity sectors:

a) Primary sector: agriculture, farming, forestry and fishing activities.

b) Secondary sector: industry and construction.

c) Tertiary sector: wholesale, retailing (food, non-food and department stores), hotels and restaurants, transport and communications, financial and real estate, education, health and social services, and professional, artistic and leisure activities.

Figure 3.4(a) represents the zoning map of Extremadura, which is an inland autonomous community of southwestern Spain whose capital city is Mérida. It is a large region, compared to Spain as a whole, with more than one million inhabitants and very low population density (26 km^2), which is divided into two provinces (NUTS 3), Badajoz and Cáceres. Located equidistantly between Madrid and Lisbon, it is a great hub to access the Spanish and Portuguese markets through its good communications with the most important Atlantic seaports of the Iberian Peninsula. However, Extremadura has traditionally been a rural impoverished region of Spain whose difficult conditions pushed many of its young people to seek their livelihood elsewhere and even overseas. For this reason, it is the only Spanish region receiving structural funds from the European Union. In spite of this secular backwardness,

TABLE 3.1: "DataSpa" functions used in the 2017 Socioeconomic Atlas of Extremadura

Atlas chapters	Statistical information	Package functions				
		Download	Loading	Manipulation	PDF extraction	Web scrapping
I: Economic indicators	Population	*getbase.pob()*	*pob.tot()*			
	Unemployment by sex	*getbase.paro()*	*paro()*	*pob.ind()*		
	Vehicle fleet by type				*parque.aut()*	
II: Demographic indicators	Population by sex and age groups	*getbase.pob()*	*pob.tot()* *pob.h.tot()* *pob.m.tot()* pob.a() *porb.q()*	*pob.ind()* *pob.ind.p()*		
	National and foreign population by sex	*getbase.pob()*	*pob.n()* *pob.e()*			
	Demographic phenomena	*getbase.fen()*	*pob.fen()*	*pob.ind()*		
IV: Trade areas	Population	*getbase.pob()*	*pob.tot()*			
V: Entrepreneurship						*data.firm()* *data.firm.a()*
VI: Evolution of indicators	Population by sex panels	*getbase.pob()*	*pob.tot()* *pob.h()* *pob.m()*	*pob.ev()* *pob.h.ev()* *pob.m.ev()*		
	National and foreign population panels	getbase.pob()	*pob.e()*	*pob.n.ev()* *pob.e.ev()*		
	Demographic panels	*getbase.fen()*	*pob.fen()*	*ind.ev()*		
	Unemployment panels	*getbase.paro()*	*paro()*	*ind.ev()*		
	Vehicle fleet panels				*parque.aut()*	
VIII: Mancomunidades	Population	*getbase.pob()*	*pob.a()*			
	Demographic phenomena	*getbase.fen()*	*pob.fen()*			
	Unemployment	*getbase.paro()*	*paro()*			
IX: Municipality maps	Population by age groups	*getbase.pob()*	*pob.q()* *pob.e()*	*pob.ind()* *pob.ind.p()*		
	Foreign population	*getbase.pob()*	*pob.e()*			
	Demographic phenomena	*getbase.fen()*	*pob.fen()*	*pob.ind()*		
	Population panels	*getbase.pob()*	*pob.tot()*	*pob.ev()*		

Source: self-elaboration

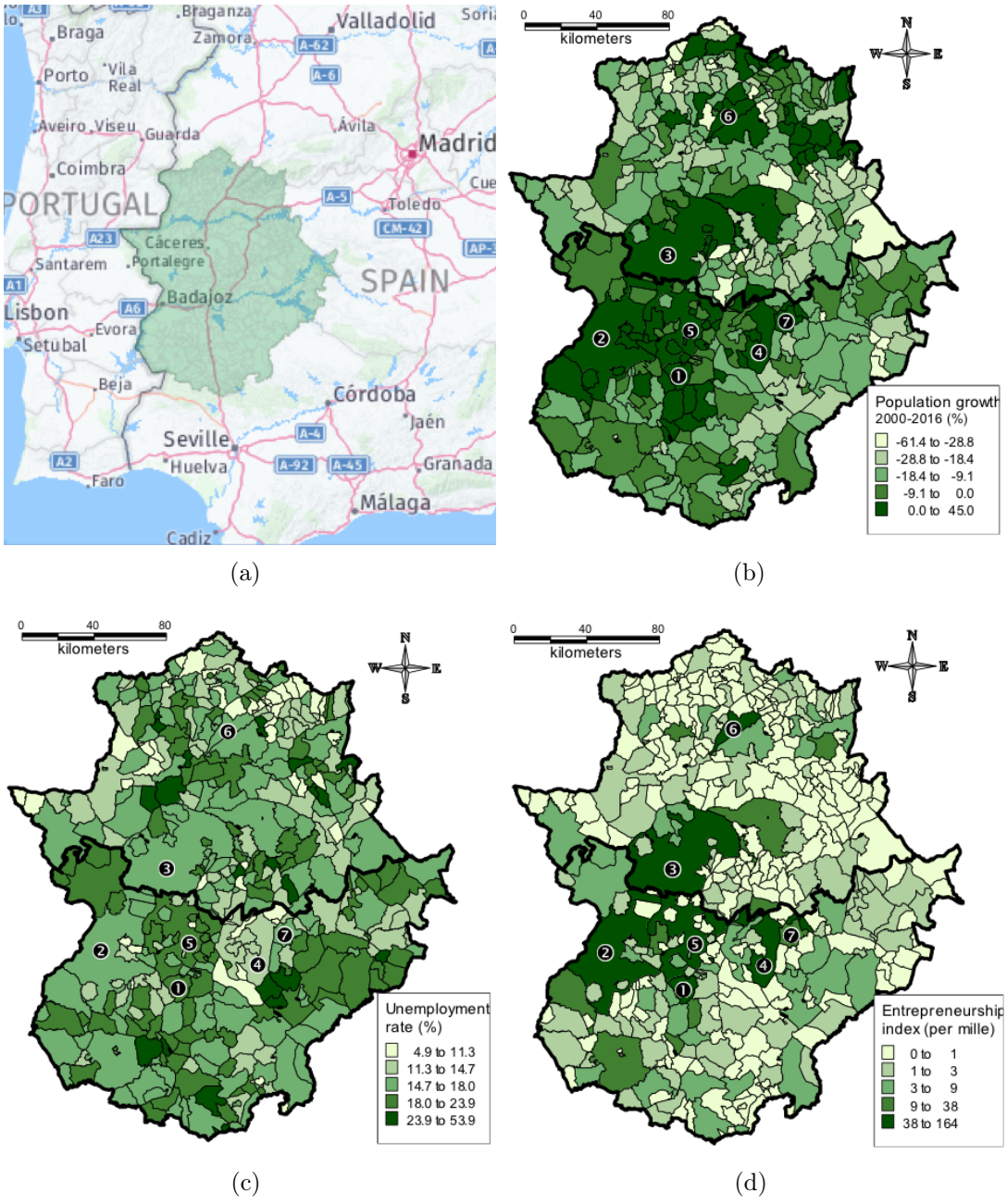


FIGURE 3.4: Zoning map of the region of Extremadura (Spain) and thematic maps of some indicators treated by the "DataSpa" package for the 2017 Socioeconomic Atlas of Extremadura. The main municipalities, with more than 25,000 inhabitants, are 1 Almendralejo, 2 Badajoz, 3 Cáceres, 4 Don Benito, 5 Mérida, 6 Plasencia and 7 Villanueva de la Serena.
Source: self-elaboration

Extremadura is currently a region where incipient network of RTD+i centers supporting entrepreneurs, wildlife, ancestral customs and historic cultural heritage come together in perfect harmony.

In order to have a better knowledge of such a diverse autonomous community, the regional government of Extremadura publishes the Socioeconomic Atlas biennially, which contains almost 200,000 data and more than 400 variables. In Figure 3.4(b) and 3.4(c), we represent two indicators extracted and built –from INE and SEPE, respectively– with the "DataSpa" URL parsing functions. As shown in 3.4(b), during the last 15 years, population growth was negative at the level of municipalities, except in the bigger towns and their surroundings. This evolution is part of the "population desertification" process existing in the inlands in Spain, which particularly affects the peripheral Extremaduran municipalities limiting the two Castiles and Andalusia. Unemployment (C) is still a big issue in this region specially affecting, among others the commonwealths of municipalities ("mancomunidades") located at two main natural reservoirs of the Tagus and Guadiana rivers. Finally, we illustrate the distribution of the entrepreneurship index 3.4(d), which is the share (in per thousands) of the local firms and self-employed –web scrapped from the Axesor database with "DataSpa"– over the regional aggregates. Entrepreneurial activity is concentrated in the main cities, though there are also some intermediate centers arising from this cores in the towns of Alburquerque and Jerez de los Caballeros (west), Zarza de Granadilla, Navalmoral de la Mata and Villanueva de la Vera (north) and Llerena (south).

3.6 Conclusions

Internet undoubtedly increases the information availability and the way that researchers interact with data. Nowadays the use of internet-based databanks increases the chance of access to a large amount of‘ primary and secondary information. However, information disposal varies depending on data category and country. Major difficulties arise with geographical downscaling. In fact, the collection of microdata at low geographical level could become a challenge for urban and intra-urban analysis.

Particularly at these lower geographical scales, researchers may deal with data availability and accessibility problems. Accessibility problems are caused when the way that data is published in the servers, blocks or delays the download process, most likely producing errors in the construction process of big databases. Availability problems usually arise when the official agencies restrict access to the information producing empty data records and incomplete databases. To overcome these problems, it is necessary the use of new data extraction strategies and explore new information sources. In this paper, we present a set of

functions, which explore different methods and sources to generate databases for Spain at the municipality level (NUTS 5).

Using the URL parsing strategy, we build a set of functions to download, load and manipulate population and unemployment databases solving the accessibility problems present in two official agency web portals. We solve the accessibility problems in the vehicle fleet database using a combination of URL parsing and PDF extraction strategies. We build the `parquet.aut()` function, which employs a URL parsing strategy to download the PDF files with the municipality reports from the DGT web portal, in order to extract the statistical data with a PDF extraction strategy. We must deal with availability problems in the construction of the firm database. For this reason, we apply a web scraping strategy with the functions `data.firm()` and `data.firm.a()` to download the information of firms and freelancers freely published by a private company. The creation of a firm database is very helpful in having a knowledge about the distribution of economic activities in Spain at urban and individual level.

All these functions compounds the "DataSpa" R package, which is freely accessible under a GPL-2 license. This package is useful as a case example for countries and regions with similar statistical problems than Spain. It allows to researchers, entrepreneurs and policy makers to have a better knowledge specifically of the Spanish cities and regions by the elaboration of statistical information systems. This is the case of the 2017 Socioeconomic Atlas of Extremadura, for which "DataSpa" was built, which constitutes the most important official database of municipality variables in this region.

4 The dynamics of patentability and collaborativeness in Chile: an analysis of social networks between 1989 and 2013[1]

4.1 Introduction

More recent studies have used patents as important indicators of the level of research activity (Furman, Kyle, Cockburn, & Henderson, 2006), collaboration between companies (Lecocq & Van Looy, 2009), and technological specialization of a country (van Zeebroeck, de la Potterie, & Han, 2006). Patent data provide fairly detailed descriptions of the most up to date and technically feasible technologies, including information on the assignees, inventor names, dates, location, and invention claims rarely found elsewhere (Jaffe & Trajtenberg, 2002). Hence, these data allow large-scale analyses (Leydesdorff, 2008), enabling researchers to perform studies of industrial trends and research collaboration over extended time periods (Huang, Chiang, & Chen, 2003).

However, in Latin America, studies on intellectual property and research collaboration are often limited by the lack of reliable empirical datasets. Different from most industrialized economies, comprehensive databases are often incomplete and researchers must collect data directly on a case-by-case basis (Barroso, Quoniam, & Pacheco, 2009). In the specific case of Chile, only few studies on patents are available, with most of them compiling their own datasets (Abud, Fink, Hall, & Helmers, 2013; Bas & Kunc, 2009; Modrego, McCann, Foster, & Olfert, 2015). Large international standardized databases such as Thomson-ISI, and USPTO (Krauskopf, Krauskopf, & Méndez, 2007), and the Innovation Survey data (Benavente, 2006; Crespi & Zuniga, 2012) have also been used by scholars to estimate innovation activity in Chile. These studies provide, however, little information on patents, at least partly, because of the lack of patenting culture in Chilean companies and universities (Krauskopf et al., 2007).

[1] A first version of this chapter has already been published in Pinto, Vallone, Honores, and González, 2017.

This might also explain the limited number of studies using patent data which examine research collaboration between companies as well as science-industry interactions in Chile. There are two recent exceptions. First, Abud et al., 2013 use co-assignment patterns in patent filings between resident and non-resident companies, and between universities and companies as a measure of innovation output. Their results show not only little evidence for collaboration in Chile, but also reveal that co-assigned or joint patents resulting from collaborative efforts account for a small share of patents in Chile—3% between 1991 and 2010—. This study, however, provides information based on the number of patent applications rather than those that are actually granted, and therefore it may distort the actual number of effective collaborations. Second, Morales Valera and Sifontes, 2014 provide some evidence on co-inventions in some Latin American countries, including Chile. Their results, however, only take into account collaboration at the individual level (inventors), but not between organizations (e.g., companies).

Our paper enters that literature with some new results, based on a patent dataset compiled from the Chilean National Institute of Industrial Property (INAPI) specifically for the purpose of this study. The aim of the paper is two-fold: the first is to provide a quantitative overview of the patenting landscape in Chile during the period of 1989 – 2013, which will serve as a background to the analysis of research collaboration given in the second part of the paper. Among other things, we look at how grant lags vary by groups (e.g., individuals, companies, and universities) and residence of assignees. Second, based on these statistical data, the paper explores the country's patterns of research collaboration.

Methodologically, social network analysis (Scott, 2013) is used for this purpose. First, we develop a search and replace algorithm in R to deal with the aforementioned data quality problems. Second, we use social network analysis to understand the structure and dynamics of collaborative research production in the country. In terms of graph theory, the dataset was initially arranged as an assignees-by-event incidence matrix M to generate a bimodal network, which allows exploring graphically the relationship between productivity and collaboration. From the product matrix of M with its transpose M', which generates an adjacency matrix and a unimodal symmetrical valued and undirected network, it is possible to study in depth the collaboration structure of the Chilean firms. In addition, we also contrast the relationship between productivity and collaboration with the Moran's I and global G autocorrelation tests, using the adjacency matrix as 'vicinity' weights.

The evidence of our results suggests an intensification of the knowledge content out of domestic and foreign companies reflected in a growing number of patents granted in Chile, but also reveals an overall small contribution of universities to the total percentage of these patents. We also found that patent grant lags were long and variable. Variation in patent grant delay seems to be driven by differences across different technological fields. Surprisingly, shorter

grant lags in Chile were found for domestic and foreign assignees, for universities over companies, and for repeat patentees over first-timers. Overall, our study points to the lack of collaboration between industrial partners as well as science-industry, at least in terms of patent co-ownership. As such, Chile's patent ecosystem looks poor and highly concentrated within groups, mainly around large non-resident companies.

The rest of the chapter is organized as follows: Section 4.2 characterizes data; Section 4.3 maps the patent ecosystem in Chile using social network analysis, and Section 4.4 discusses and concludes the study and our findings.

4.2 Patent data

The empirical analysis is based on a patent dataset comprising all national and foreign invention patents that have been effectively granted in Chile, all of which are still valid in 2014. Patents are identified using INAPI's own identification code. The file contains data ranging from the patent application date (or filing date) to the date when the patent was granted (or entitlement date), and from the full names of each of the assignees and inventors listed in each patent to the patent description and the different technology classes to which each patent pertains. The latter is registered according to the International Patent Classification (IPC)[2]. The dataset also contains data on resident, multi-resident, and non-resident assignees. We classify the assignee of the patent as resident or non-resident (i.e., individuals or entities that do not have Chile as their country of primary residence), based on its address. Multi-residents are those that file for patents sometimes through an entity resident in Chile and some others through an entity abroad.

The initial database of patents in Chile contained more than 66,500 observations filed from June 19, 1985 through June 7, 2013 (about 29 years). Due to severe inconsistencies, we had to eliminate those patents filed before 1989 leading to a total sum of 66,536 patents. These data has also been reduced progressively in a cleaning and selection process that is explained hereafter.

4.2.1 Cleaning the database with the 'msp' package

Raw data presents data accuracy problems, discrepancies such as spelling errors, acronyms, abbreviations, and names listed differently (e.g., "Inco Limited" and "Inco Ltd."). Different assignee names referring to the same individual or entity (e.g., "Pontificia Universidad

[2]We use IPC version 8, which includes the following sections: A. Human Necessities, B. Performing Operations; Transporting, C. Chemistry; Metallurgy, D. Textiles; Paper, E. Fixed Constructions, F. Mechanical Engineering; Lighting; Heating; Weapons; Blasting, G. Physics, and H. Electricity.

Catolica de Chile Facultad de Ciencias Biologicas" and "Pontificia Universidad Catolica de Chile" were both referred in the database as to "Universidad Catolica de Chile (PUC)"). Assignee names changing over time were also tracked and modified accordingly. To avoid duplications, each assignee's name was harmonized by the use of an interactive search and replace algorithm built in R, 'msp' [3], which is shown in Figure 4.1.

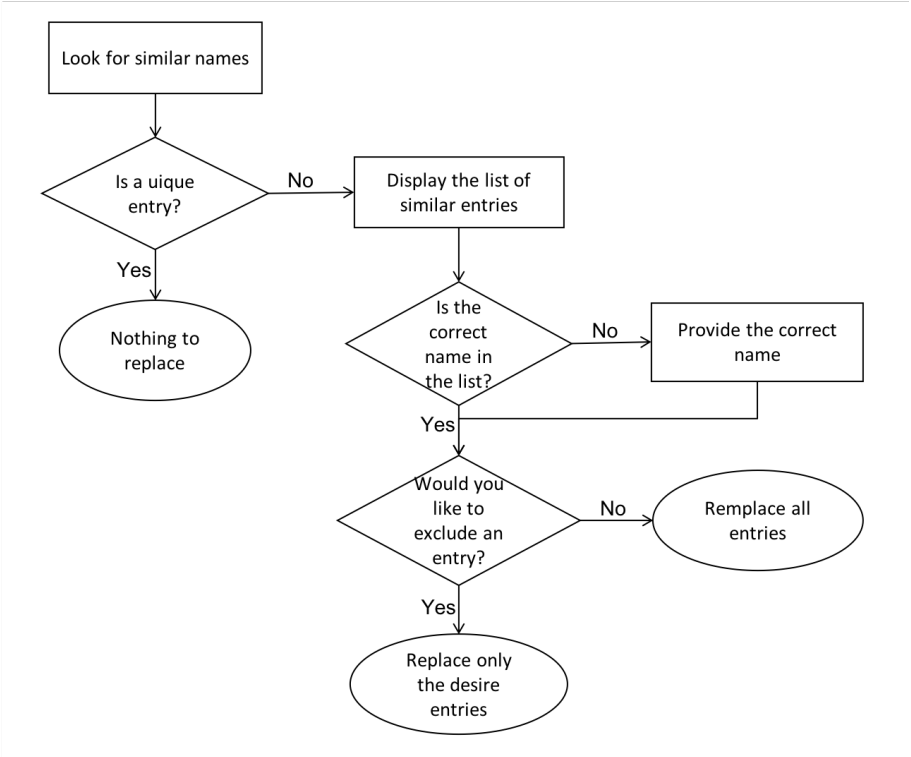


FIGURE 4.1: "msp" algorithm process
Source: self-elaboration

In the first step, the algorithm looks for similar values of an entry based on the Jaro–Winkler distance (Winkler, 1990). In case of not finding any similar entry, this entry is considered a unique entry and it is not necessary to make any replacement. When similar entries are found, the algorithm provides a list of the similar values found and ask the user to choose the correct value to replace. The user must provide a valid name. Two valid entries could be considered as similar by the algorithm (e.g José Morales and Joan Morales). In this case, it is possible to exclude them from the replacement list.

When there is of more than one assignee to the same patent, we separate the names since the content is all contained in one cell. Our revision of the raw data also revealed inconsistencies between application and granting dates. We checked and (when needed) corrected the data using two other databases, a web-based search tool[4] , and an INAPI proprietary patent

[3]This R package is freely available from the site https://github.com/amvallone/msp. It must be installed in the R console with the command: `devtools::install_github("amvallone/msp")`.

[4]http://ion.inapi.cl:8080/Patente/ConsultaAvanzadaPatentes.aspx

collection database. We opt to remove from the database other important dates—such as publication, registration, and expiration—due to the lack of reliable information. To avoid duplications, the final database uses the INAPI's unique patent identifier number for each of the listed patents.

4.2.2 Granted patents

In this study, we use data based on patents granted rather than on patent applications to avoid overestimations of the invention numbers. In effect, Archambault, 2002 argues that patent applications cannot be counted as new, useful or non-obvious, until they are granted. Hence, using data on granted patents are becoming more frequent (e.g. Dosi, Grazzi, and Moschella, 2017), although this choice restricts the size of the dataset. In fact, this selection limited our initial database of 66,536 patents to 42,313 granted patents registered from 1989 through 2013.

Additionally, it must be said that since the Chilean law allows applicants to register a patent in different IPC sectors, classes and subclasses, depending on what the final product is expected to being used for, the same patent could have more than one entry in the database. In fact, the average number of entries per patent was about four. After removing duplicate observations, we ended up with 9,938 granted non-duplicate patents in our dataset.

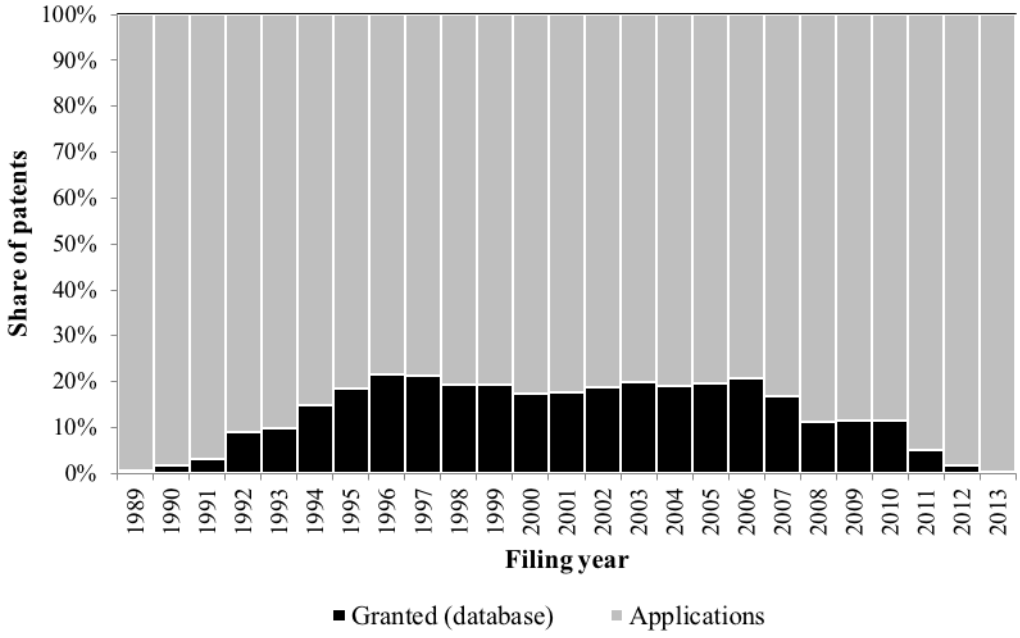


FIGURE 4.2: Ratio of active patents granted to patent applications at INAPI, 1989–2013
Source: self-elaboration from INAPI; and INAPI Biblioteca Digital.

Hence, as shown in Figure 4.2, the average value reported in the INAPI Bulletin series in the period 1989–2013 is 17.6 percent for patents granted relative to those applied in Chile (INAPI, 2016). This seems to be a reasonable number, consistent with the information available. Considering the percentage of patents granted relative to those applied in Chile, our results are similar to those of INAPI, 2016 and Hernandez-Cuevas and Valenzuela, 2004. Thus, we believe that we have adequately covered the entire period 1989–2013.

It might be argued that Chile is a developing country and as such, the low success rate of about 17% for patent applications (number of granted patents / number of patent applications) might be due to its underdeveloped IP system. However, the situation in Chile is worse when compared to other countries in the region with similar GDP per capita levels, like Brazil (48%) and Argentina (24%) (Hernandez-Cuevas & Valenzuela, 2004): these other countries have gone ahead regardless of their political instability or incipient institutional system. This outcome for Chile can also be considered to be inefficient in comparison with American developed countries like the US (62%) and Mexico (50%) (**Archambault:2002fz**).

4.2.3 Main characeristics and limitations of the final database of granted patents

The size of our dataset is influenced by other two factors. First, Chile provided 15 years of patent protection prior to the WTO agreement entering into force (Law 19.996 of 2005). Thus, granted patents, which were filed prior to year 1990, are unlikely to appear in our dataset—in fact, they correspond to less than 0.05

This is in line with previous findings within different Chilean samples (Abud, Hall, & Helmers, 2015; Schmal, López, & Cabrales, 2006). This is also why less than 3% of the data in our file corresponds to patents filed after 2010. Thus, considering lags in patenting, decreasing records in our database towards the end of the period are influenced by truncation of data as a larger share of patents still awaits the examination decision.

With respect to the origin of the patent, data show significant differences between domestic and foreign assignees, and between countries. These differences have also changed over time. For instance, granted patents to domestic assignees have grown from 5.9% to 12.7% over the past decade and a half, whereas countries like the U.S., and Germany have registered a relative drop in participation in the total number of granted patents in Chile. Overall, however, the number of patents for the studied period has grown, for both domestic and foreign assignees, which might suggest an intensification of the knowledge content of their businesses in Chile (Figure 4.4). The growth in the number of patents granted to domestic assignees in Chile can be a result of two combining factors: a growth in the country's

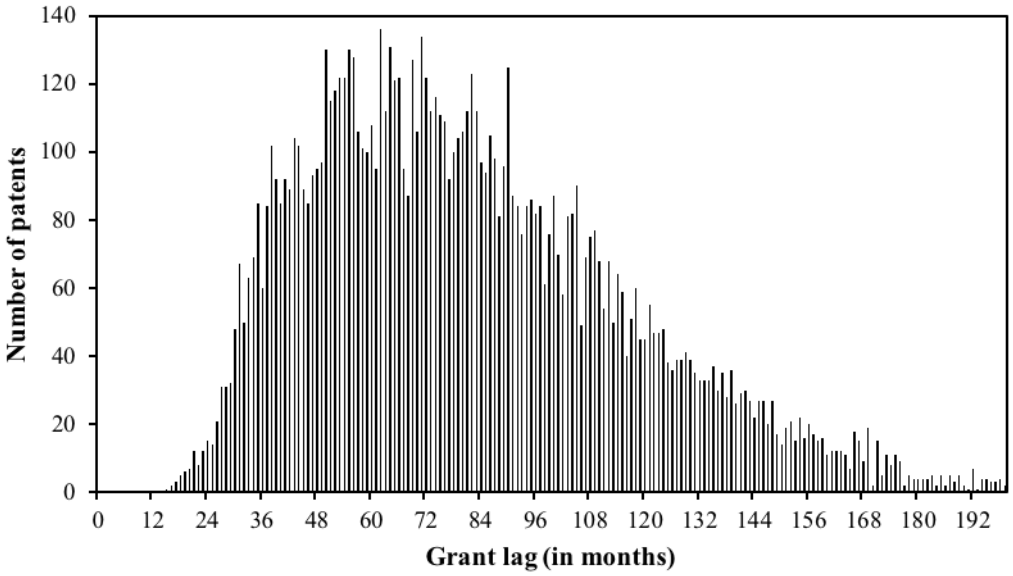


FIGURE 4.3: Distribution of grant lags in Chile, 1989–2013*
Source: self-elaboration
* The distribution of grant lags covers 99.5% of all patents. The actual tail of the distribution continues beyond what is shown in the graph (266 months long).

patenting activity (i.e., more patents granted to local residents), and also a change in the countrys residence of patentees, from non-resident to local. The latter is often associated with multi-resident entities that have both local and international addresses.

Also in terms of assignees, data show that the top-20 assignees in Chile are all non-resident companies. They represent 27% of the total granted patents in Chile for the period. Companies like Unilever (382 patents), Nestlé (220), and Bayer (218) lead the charts. Among Chilean residents, the top-20 resident assignees represent 33% of the total number of patents granted to Chilean assignees, although they correspond to only 3.3% of total patents in Chile. Three of the top-20 resident patent assignees (Instituto de Innovacion en Mineria y Metalurgia, Corporacion Nacional del Cobre de Chile, and Biosigma) belong to the Chilean state owned copper mining company, CODELCO (Table 4.1). An element to consider is that in the past 20 years, 45% of the total patents granted in Chile to the top-20 assignees relates to mining.

The second tier of top-patent resident assignees is university (40%). IP co-ownerships between universities and other entities are, however, relatively rare. For instance, among the top five ranked universities in terms of granted patents, less than 5 percent are co-owned.

Regardless their residency, companies by far dominate the landscape (92%), followed by individuals (6%). The remaining 2% of the granted patents are universities. This indicates a relatively small overall contribution of universities in Chile in the generation of new,

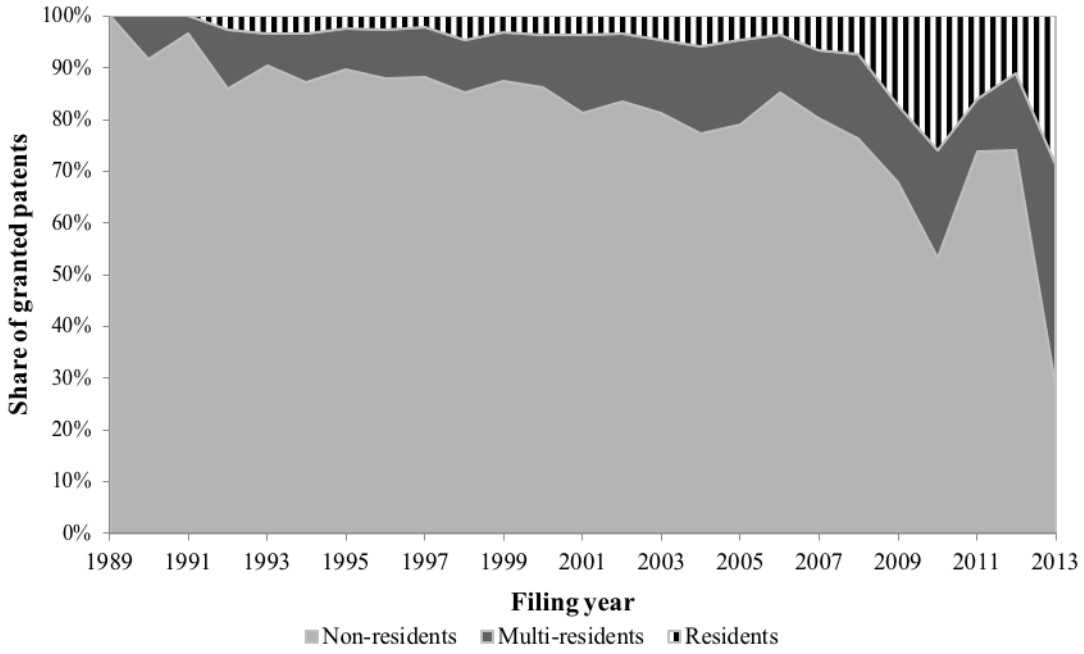


FIGURE 4.4: Share of granted patents by assignees' residence type and filing date
Source: self-elaboration

patentable knowledge. However, the low percentage (1-2%) of university patenting observed in the US Leydesdorff:2010he reflects a global rather than a local phenomenon that requires more systematic investigation.

TABLE 4.1: Chile's top-20 resident assignees

Rank	Assignee (patentee) name	Patents	% Total	Industry
24	Instituto de Innovacion en Mineria y Metalurgia S.A.	50	0.49%	Basic Materials*
26	Corporacion Nacional Del Cobre De Chile	46	0.45%	Basic Materials*
37	Universidad Tecnica Federico Santa Maria	35	0.35%	University
41	Universidad De Concepcion	34	0.34%	University
66	Universidad Catolica De Chile (PUC)	22	0.22%	University
93	Universidad De Chile	16	0.16%	University
102	Compania Chilena De Tabacos S.A.	15	0.15%	Manufacturing
114	Universidad De Santiago De Chile	14	0.14%	University
115	Vulco S.A.	14	0.14%	Basic Materials*
137	Cintac S.A.	11	0.11%	Construction
141	Minera Michilla S.A.	11	0.11%	Basic Materials*
165	Fundacion Chile	9	0.09%	Technology
194	New Tech Copper S.A.	8	0.08%	Basic Materials*
177	SQM	8	0.08%	Basic Materials*
204	Biosigma S.A.	7	0.07%	Basic Materials*
205	Cartones San Fernando Ltda.	7	0.07%	Manufacturing
214	Instituto De Investigaciones Agropecuarios (INIA)	7	0.07%	Agriculture
221	Industrias Metalurgicas Sorena S.A.	7	0.07%	Manufacturing*
221	Universidad Catolica De Valparaiso	7	0.07%	University
221	Universidad Catolica Del Norte	7	0.07%	University
	Total	335	3.31%	

Source: self-elaboration

* Patents can be related to the mining sector.

Surprisingly, though, there are significant benefits to being a resident as compared to being a multi-resident or a non-resident, and to being an individual or university as compared to being a company in terms of length of time to obtain a patent in Chile. For instance, a one-way ANOVA comparing grant lag means for resident, multi-resident, and non-resident groups has revealed statistically significant differences between groups ($F(2, 10134) = 100.1$, $p < 0.000$). Multiple comparisons showed that all three groups differed significantly from each other group according to Scheffe's test for all pairwise comparisons at a significance level of 1%. Among all three groups, non-resident assignees show longer grant lags (an average of 7.1

years, where for residents and multi-residents is 5.3 and 6.4 years, respectively), which may suggest that being a foreigner is a liability (Table 4.2). Previous studies in other contexts (e.g., the U.S.) have suggested that applications may take longer for communication and logistical reasons (Popp, Juhl, & Johnson, 2004), although this point needs more systematic study in the case of Chile.

There are also statistically significant differences between assignee group means as determined by one-way ANOVA (F(2, 10134) = 27.99, p <0.000 (p-value <0.01)). Multiple comparisons show that university and individual groups differed significantly from company groups according to Scheffe's test for all pairwise comparisons at a 99% of significance (p <0.01). In Chile, the grant lag for companies is on average 7 years, where for individuals and universities is 6.2 and 6.3 years, respectively.

TABLE 4.2: Grant lags (in years) by residency type

Overall Descriptive Statistics	Residents	Muti-residents	Non-residents	Company	University	Individual
Frequency	494	1,249	8,391	9,291	224	618
Mean	5.3	6.4	7.1	7	6.3	6.2
Std. Dev.	2.6	2.8	3	3	2.8	3.2
Median	4.8	6.1	6.7	6.5	6.1	5.5
Minimum	1.5	1.4	1.3	1.3	1.8	1.5
Maximum	18.5	19.9	22.2	22.2	15.2	18.5

Source: self-elaboration

4.3 Network analysis

We use social network analysis to map Chile's patent ecosystem. Using data from patent co-ownership, we reveal the country's underlying research collaboration patterns, and the productivity levels of the individuals and entities involved in the generation of new patents. Assignees with two or more patents were considered.

Research collaboration is assumed here for patents that are jointly owned by two or more unrelated assignees. As the literature suggests, the analysis of co-ownership provides a basis for understanding collaboration between companies as well as science-industry (Lecocq & Van Looy, 2009). Since we have information on all assignees, we limit co-ownership to related or unrelated entities, and therefore we avoid counting a patent that is co-owned between partners belonging to the same company (e.g., collaborations between headquarters and affiliates, or between privately owned research centers and their parent companies, like Nestlé Research Center). By the same token, research collaborations across university

departments were excluded from the sample. This allows us to actually capture and map patent collaboration between distinct, unrelated domestic and foreign entities (Abud et al., 2013). As indicated above, no analysis on inventor affiliation was performed.

Methodologically, we perform our analysis by building a network of affiliations (Scott, 2013) where assignees or patent owners are actors and patents are events (check in Annex 1 the R code used for computations). Following Borgatti and Halgin, 2014, the term "affiliations" refers here to co-ownership (partner entities or individuals having participated in a focal patent), and the assumption is that co-ownership is an indicator of an underlying social tie.

The network construction is based on a incidence matrix M, which assumes the value 1 if the assignee is a patent holder and 0 otherwise. By the row sums of matrix M it is possible to measure the assignee productivity and from the sums of the M matrix columns it is possible to obtain the level of collaboration of each patent. Since we were interested in comparing collaboration (i.e., patents containing two or more assignees) with productivity (i.e., assignees with two or more patents), we remove from the incidence matrix all the assignees with a single non-collaborative patent. In practice, the removing process has two steps: first we set to 0 each m_{ij} which jointly fulfil the condition $\sum_i m_{ij} = 1$ and $\sum_j m_{ij} = 1$ and second, we remove from matrix M all the entries $\sum_i m_{ij} = 0$ or $\sum_j m_{ij} = 1$.

The resulting network contains 8,095 patents and 1,052 assignees. To study in depth the collaboration structure of the network, we build and adjacency matrix from the product of the incidence matrix M with its transpose, $A = MM'$, leading to a co-affiliation network (Borgatti & Halgin, 2014). The main diagonal of matrix A contains the patent productivity of each inventor, that is, the number of patents that, individually or collaboratively, is produced by each assignee. Cells out of the main diagonal of matrix A show the number of patents collaboratively produced by two different assignees. Since the main objective of the co-affiliation network is to focus in collaborative analysis, we delete all the isolated nodes from the network; that is all the rows and columns with positive values only on the main diagonal of the adjacency matrix. Hence, we obtain a reduced network containing 222 assignees who only produced 166 patents in total.

In addition, we want to explore the existence of productivity spillovers into the co-affiliation network. Network spillovers take place when on the one hand, productive assignees tend to be related to each other leading to a 'productive cluster' and, on the other hand, non-productive assignees also tend to have a connection among themselves forming a 'non-productive' cluster. From the adjacency matrix $A = \{a_{ij}\}$ we generate a vicinity weights matrix $W = \{w_{ij}\}$, which is as a binary matrix such that $w_{ij} = 1$ for $a_{ij} > 0$ and 0, otherwise, setting to zero the main diagonal. This is a matrix of weights, which allows us to test the existence of a relationship between productivity and collaboration using the Moran's I (Anselin, 2013) and global G (Getis & Ord, 1992) autocorrelation tests.

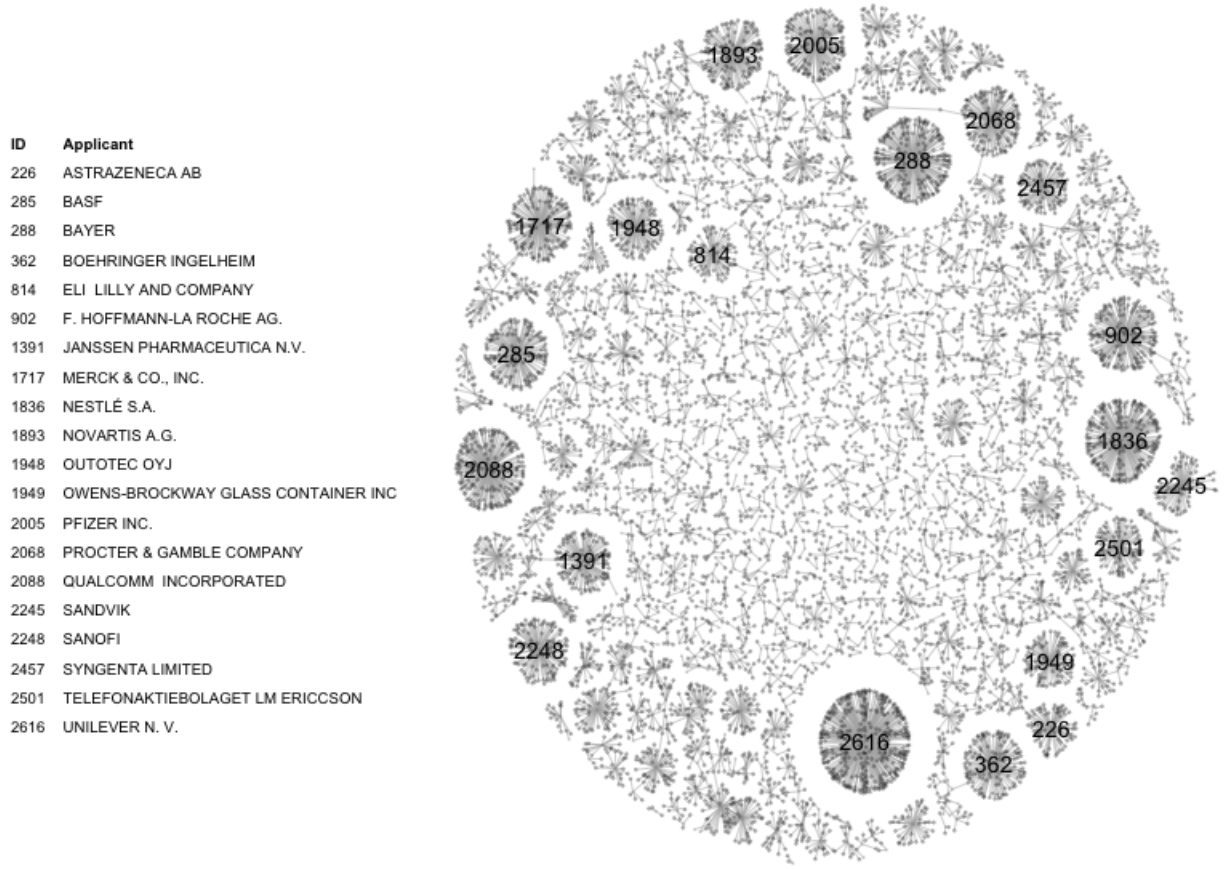


FIGURE 4.5: Granted patents and assignees network graph (Fruchterman-Reingold)*

Source: self-elaboration

* Gray squares correspond to assignees and black circles to patents.

In Figure 4.5, we plotted the results for assignees (gray squares) and patents (black circles) using the Fruchterman-Reingold algorithm (Fruchterman & Reingold, 1991) in order to avoid cutting ties and to reveal the clusters more clearly. Network analysis reveals several isolated groups, suggesting that joint patenting is a rare event. This is in line with findings in previous studies in Chile (Abud et al., 2013) and also in the US (Kim and Song, 2007). In addition, the data suggest that the most highly productive patentees, those with 60 or more patents, had little or no ties with others in Chile. When one computes the figures over the top 20 most productive companies in Chile—which includes, among others, multinational consumer goods companies like Unilever, Nestlé, and Procter Gamble, and global pharmaceutical companies like Bayer, Hoffmann-La Roche, Merck, Pfizer, and Novartis—, only twenty of their patents (0.7%) were jointly taken. These companies are identified with a code number in Figure 4.5.

In Figure 4.6, a co-affiliation network is plotted. The results yield a simple network, with

small isolated groups and associations of two or three assignees standing out. The degree centralization measure is 0.055 showing that the network is not strongly connected.

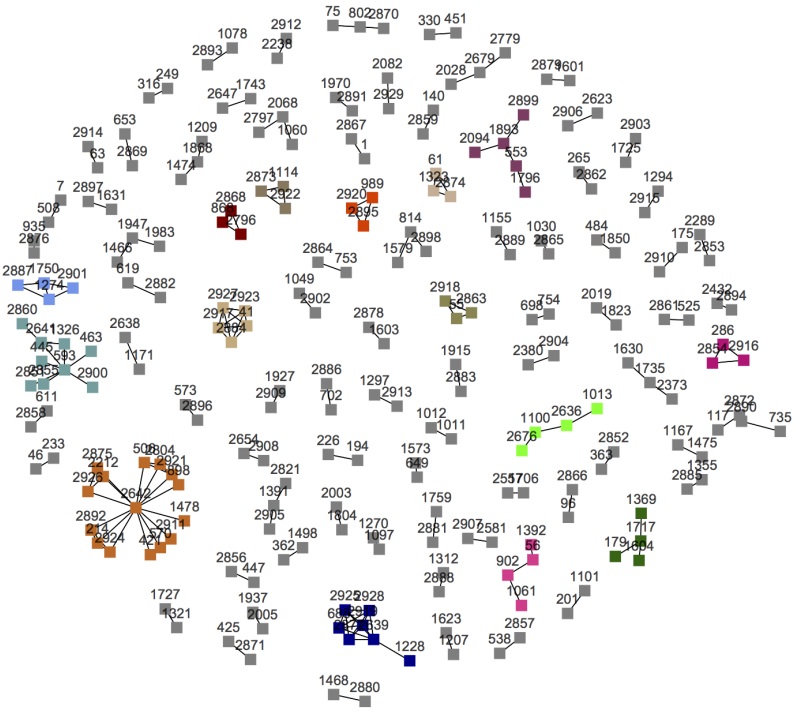


FIGURE 4.6: An assignee network (Fruchterman-Reingold)
Source: self-elaboration

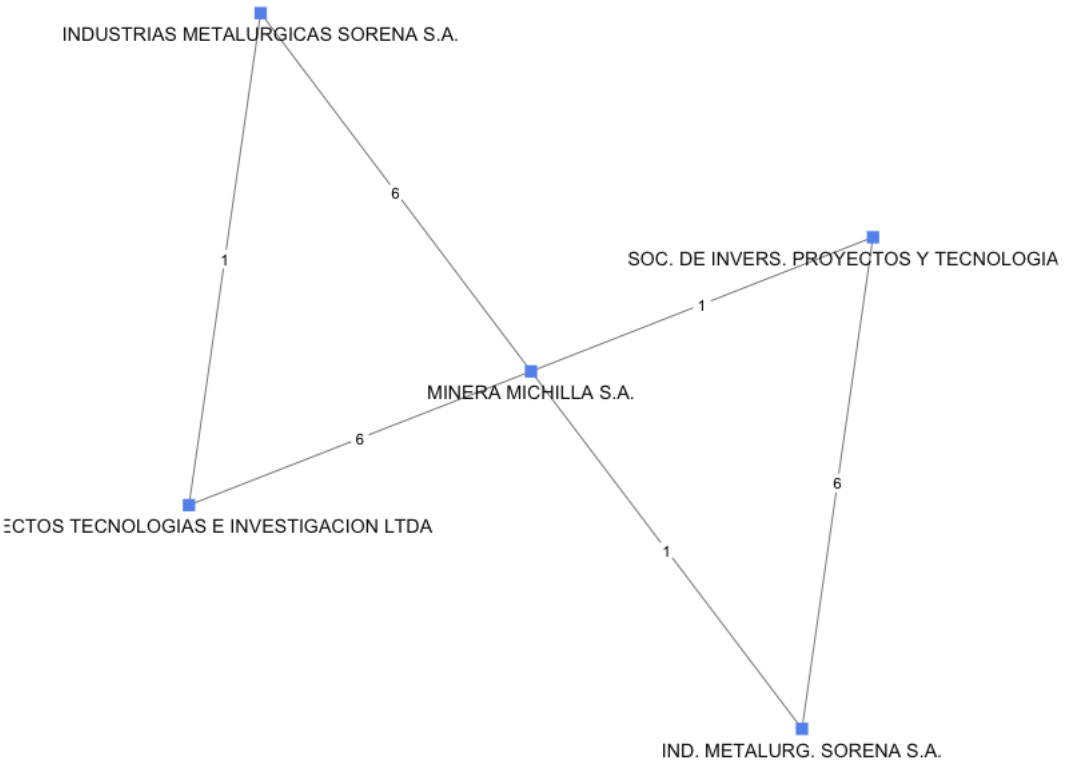
Specific results for companies with geographical ties to Chile are more illustrative. In Figures 4.7 and 4.8, we have represented the networks depicted in cyan, light blue, brown and dark blue in Figure 4.6.

On average, slightly less than 10% of all granted patents to the top 20 most productive resident and multi-resident patentees in Chile comes from joint research production. Of these, 39% (34/86) and 28% (24/86) are from the state-owned copper giant CODELCO ('Corporacion Nacional del Cobre de Chile'), and the 'Centro de Investigacion en Mineria y Metalurgia', respectively. The latter has been all jointly taken with CODELCO, which indicates a closed and highly specialized network (Figure 4.7(a)). We use code numbers to identify individuals to preserve anonymity.

Figure 4.7(b) shows one of the few networks that connects companies with companies: 'Minera Michilla S.A.', a Chilean mining company, and its technological providers.

In addition, a remarkable example of collaboration belongs to a university group member, The University of Concepcion ('Universidad de Concepcion'), with ten out of its 34 patents developed in association with others (Figure **??**). However, this university, like most of

(a)



(b)

FIGURE 4.7: Selected assignee networks of companies
Source: self-elaboration

them in Chile, shares the ownership of a patent with the faculty members who have invented the new technology. This means that the collaboration metric used in our analysis for universities should be viewed as an overestimation of the amount of collaboration, with actual levels of interactions with outsiders being lower. In fact, the UdeC has joint patents with two companies (El Indio Minery Comp. and Exxon Mobil) and the remaining are all shared with individuals (Figure 4.7(b)). The exactly same behavior is displayed by the 'Universidad Catolica del Norte' (Figure 4.8(b)).

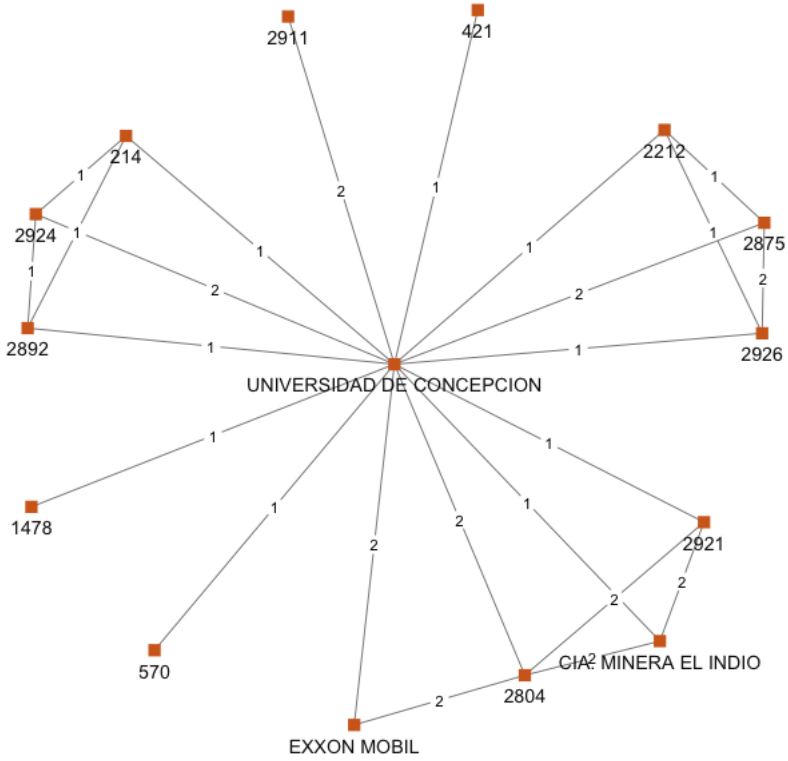
Transforming the adjacency matrix A into a vicinity weight matrix W, as explained before, we can estimate the Moran's I and global G index. The Moran's I under 1000 permutation takes the value 0.015604 with p-value of 0.2925 and the global G index takes the value 1.2747 with p-value 0.1012. Both indexes show a positive relationship between productivity and collaboration, however neither of them exhibit a strong statistical significance. Hence, we cannot clearly reject the null hypothesis of absence of productivity cluster; that is, in Chile the most productive assignees are not related each other, and vice versa.

In Figure 4.9, we depict the network using the four quadrants of the Moran's scatterplot, classifying the assignees into four mutually and exclusive categories High-High (HH), Low-High (LH), Low-Low (LL) and High-Low (HL). For example, the HH quadrant contains those assignees with a productivity level above the country average that are connected to assignees with a mean productivity level also above the average.

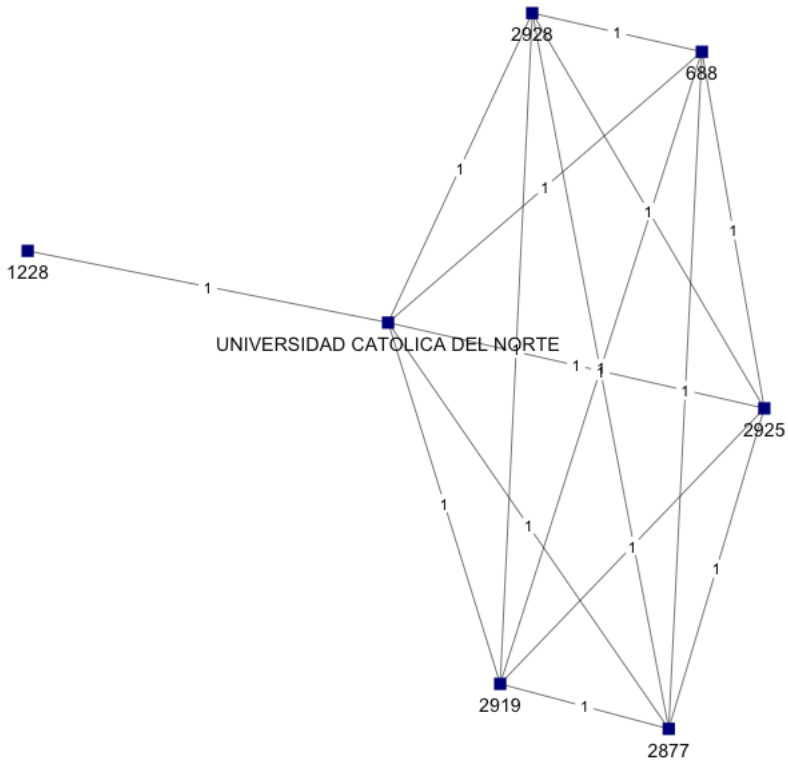
The plot shows the absence of an HH cluster in the network, since none of the clusters is entirely composed by assignees located in the HH quadrant. However, there is a predominant type of relationship in the network, the one of assignees in cluster LL (assignees with a productivity level below the country average that are connected to assignees with a mean productivity level also below the average), which denotes the existence of a low productivity cluster in the collaboration network.

4.4 Discussion and Conclusions

This paper examines patenting activity and research collaboration in Chile using a novel dataset collected from the Chilean National Institute of Industrial Property between 1989 and 2013. In terms of geographic scope, Chile is attractive as a case study due to both the country's exceptional economic growth, political stability, and international economic integration(Abud et al., 2013; Lederman, Messina, Pienknagura, & Rigolini, 2013), and the explicit experimentation with innovation policies over the past 20 years (Giuliani, Morrison, Pietrobelli, & Rabellotti, 2010; Guimon, Klerkx, & de Saint Pierre, 2016).



(a)



(b)

FIGURE 4.8: Selected assignee networks of companies
Source: self-elaboration

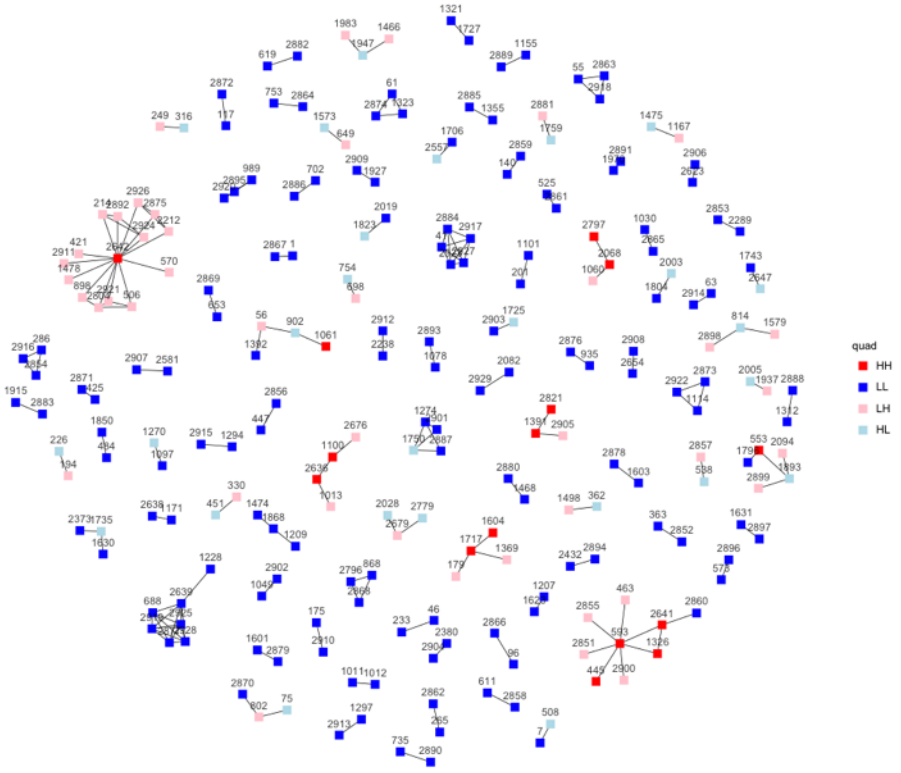


FIGURE 4.9: Productivity clusters determined by the Moran's scatterplot quadrants*

Source: self-elaboration

* HH is High-High quadrant, LL is Low-Low, HL is High-Low and LH is Low-High

The evidence reveals an intensification of the knowledge content out of domestic and foreign companies reflected in a growing number of patents granted in the country, but also an overall small contribution of universities to the total percentage of such patents. The results of our study also reveal that patent grant lags are long and variable.

Overall, the results of our study reveal a lack of collaboration between companies and also between science and industry. The preponderance of companies and universities without any ties at all in Chile is clear, and, as such, the country's patent ecosystem looks poor and extremely concentrated within groups, mainly around large non-resident companies. Some hope is found, however, when we explore collaborations between companies with geographical ties to Chile. This may suggest that programs that offer incentives to large international corporations to establish formal collaboration agreements with domestic universities and companies to conduct research, such as the ICE program, can help improve the health of Chile's innovation ecosystem.

We also recognize certain limitations that can lead to follow-up studies. First, the number of patents in our database represents a relatively small subset—17.6%— of the total number

of patents filed at INAPI in Chile during the period. It includes patents for which the inventions fulfil the criteria of novelty, usefulness, and non-obviousness. The study carried out was also limited to patents that were valid to 2014, which limited the presence of old data. Our results were further limited by truncation of data, which also limited the presence of new data. Future research could use more extensive data coverage to remedy this problem.

In addition, it should be noted that co-patenting indicators such as joint patent ownership do not capture all collaborations, in part due to the IP arrangements that partners might have arrived to prior to filing a patent (Lecocq & Van Looy, 2009). It is also true that joint patenting is a rare event even in developed economies (Kim and Song, 2007). Future studies could sample granted patents and ask companies, universities, and individual inventors whether they have collaborated with others in those patents in order to find traces of research collaboration across them, or alternatively study co-inventions—a situation in which a focal patent lists multiple inventors—as channels of knowledge flows.

Finally, patent data are an imperfect indicator of the number of technological innovation, especially in developing countries(Crespi & Zuniga, 2012). Some companies rely, for instance, on trade secrets rather than on patent protection to avoid leakage to the public during the application process or because the technology may be obsolete when the patent is finally issued (Bhattacharya & Guriev, 2006). Therefore, future research could use other indicators of innovation to confirm and enrich our findings.

5 Spatio-temporal methods for the analysis of the Chilean urban system dynamics

5.1 Introduction

Urbanization and economic development are closely related, since urbanization occurs as countries shift from rural-agricultural activity to urban-industrial activity (Davis & Henderson, 2003). Yet the urbanization process is not spatially homogenous. Factors such as geography (Henderson, Shalizi, & Venables, 2001), factor endowments (Venables, 2005), spatial proximity among human settlements (Ioannides & Overman, 2004), and public policies (Davis & Henderson, 2003; Desmet & Henderson, 2015) affect the evolution of urban systems, generating regionally concentrated urban settlements (Antrop, 2004). In Latin American countries, including Chile, population and economic activities are especially concentrated in the capital city and its corresponding metropolitan areas (Rodríguez, 2007), leading in many cases to overconcentration of the population and economic activity in these areas. People in Chile's provinces are fond of saying, "God is everywhere, but his office is in Santiago." In fact, the literature tends to focus on this megalopolis, forgetting the rest of the country. With the exception of Escolano Utrilla, Ortiz Véliz, and Moreno Mora, 2007, who analyze the entire group of Chilean cities, studies on this topic focus on urban dynamics of either the Metropolitan Region (MR) of Santiago (Rodríguez, González, Ojeda, Jiménez, & Stang, 2009) or other individual cities (e.g. Bustos Validiva, 2013; Escolano Utrilla and Ortiz Véliz, 2004; Santiago, Raggi, and Erices, 2016). To the extent of our knowledge, no research has quantified the effect of space on Chilean urban dynamics over the last century.

The main contribution of this paper is to improve knowledge of the Chilean urban system through a set of novel tools that enable evaluation of the influence of spatial proximity among human settlements on the evolution of the cities to detect regional differences in their spatiotemporal dynamics. With the exception of the kernel density functions and the standard Markov chain, these methods have been applied primarily to study the evolution of spatial systems for crime and income distributions. In the urban literature in particular, the

spatial Markov chain (SMC) method has been used to analyze the historical development of Spanish cities (Le Gallo & Chasco, 2008) and of Phoenix, Arizona

To detect different trends and spatial clusters, we focus specifically on how spatial proximity in the development of Chilean cities over the period 1930-2002 affects relative sizes and rankings. We seek to answer the following questions: First, are certain urban processes, such as urban sprawl and population convergence, homogeneous across the entire Chilean city system? Second, does Duranton's hypothesis (Duranton, 2016) that agglomeration effects are stronger in less-advanced countries apply to Chile? Third, to what extent will a city population grow faster or slower depending on its neighbors' growth speed?

To answer these questions, we first analyze the cross-sectional distribution of urban population by means of standard statistical analysis and nonparametric estimations of density functions for a series of years, a method proposed by Quah, 1996 and followed by many other authors (e.g. Xiufang et al., 2015; Xu and Zhu, 2009). Second, the growth process is modeled as a first-order stationary Markov chain and the role of geographical space in the transition probabilities evaluated with a set of methods based on a spatial version of the standard Markov chain (Rey, 2001). Third, we perform in-depth analysis to detect spatial regimes in the movement direction and ranking mobility of the Chilean urban distribution. The LISA Markov (Rey & Janikas, 2006) and directional LISA (Rey et al., 2011) approaches capture co-evolution of a spatial unit with its neighbors, identifying different spatial regimes in the ranking mobility of the urban distribution. We also study the existence of spatial differences in the city's growth pattern using the Global Indicator of Mobility Association (GIMA) (Rey, 2016). Finally, we determine the Ranking decomposition (Rey, 2004) as a cohesion measure that detects synchronic rank movements among spatial regimes.

To analyze Chilean urban dynamics, we selected the sample of contemporary cities from Chile's decennial censuses of 1930, 1940, 1952, 1960, 1970, 1982, 1992, 2002, and 2017 to obtain the probability distribution functions of the population of the municipalities with fewer than 5,000 inhabitants. The computations can be obtained with 'estdaR', an R package available under a GPL-2 license from the site https://github.com/amvallone/estdaR.[1]

The chapter is organized as follows. In Section 5.2, we present the database. In section 5.3, we present a summary of the 'èstadR" oackage. In Section 5.4, we develop the methodology and present the main results obtained for the Chilean urban system. The chapter ends with a concluding section.

[1]estdaR must be installed in the R console with the command: `devtools::install_github("amvallone/estdaR")`.

5.2 The evolution of the Chilean urban system between 1930 and 2002

5.2.1 Database

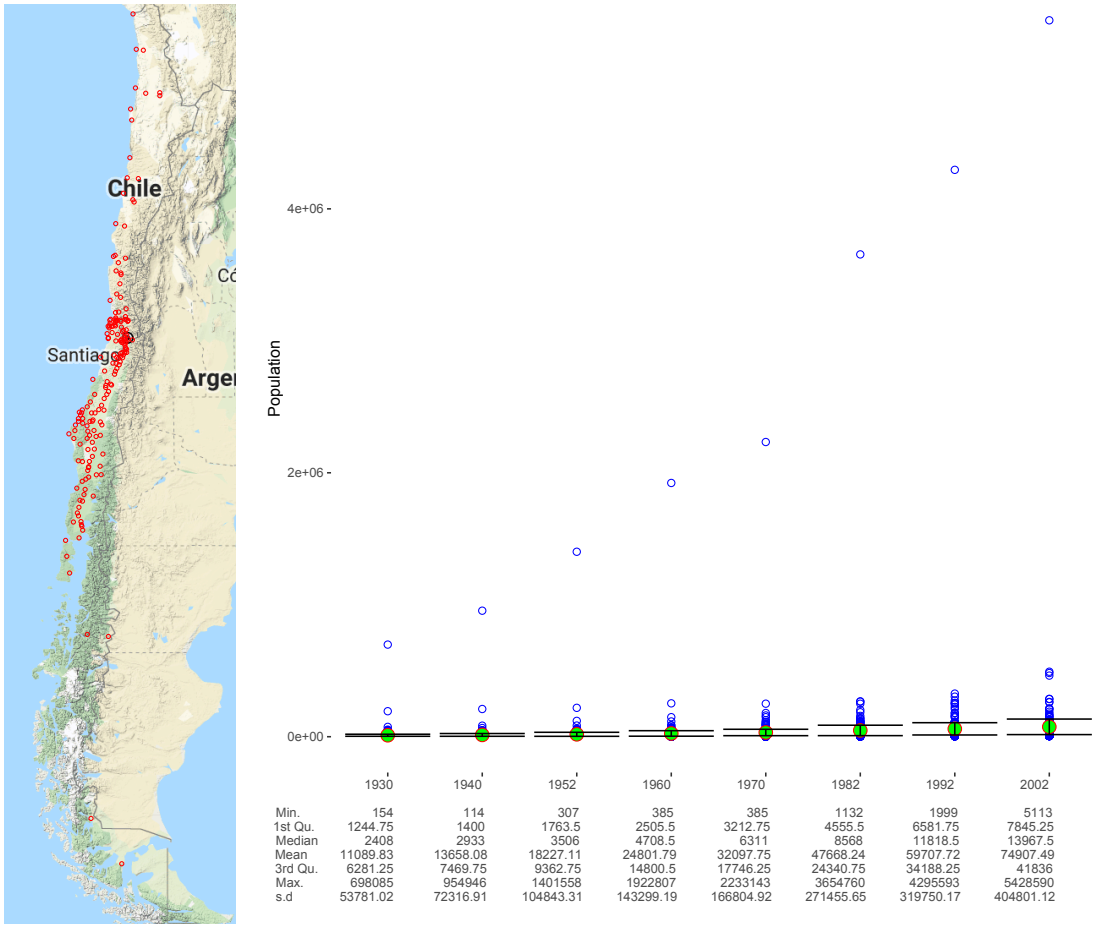
To explore the spatio-temporal dynamics of the Chilean cities from 1930 to 2002, we need a data set with urban areas defined consistently over this period. For that purpose, we have considered the definition proposed by the Chilean Office for Statistics (Instituto Nacional de Estadísticas, 1999, 2005), which considers as a 'city' any urban agglomeration with more than 5,000 inhabitants.[2] The evolution of the population distribution is analyzed using the Census data over the eight decades under consideration: 1930, 1940, 1952, 1960, 1970, 1982, 1992 and 2002.[3] The data on population are extracted from the Chilean Office for Statistics (INE) databank. Hence, we have identified the 184 Chilean municipalities over 5,000 inhabitants in the 2002 Census to study their evolution during the eight considered periods (more than 70 years). These present-day cities are located across the whole Chilean territory (Figure 5.1(a)): Tarapacá (2), Antofagasta (7), Atacama (7), Coquimbo (11), Valparaíso (30), O'Higgins (19), Maule (13), Biobío (30), Araucanía (18), Los Lagos (15), Aysén (2), Magallanes (2), Metropolitan Region (22), Los Ríos (5) and Arica y Parinacota (1).

Figure 5.1(b) presents the summary statistics for each period. The maximum values correspond to the capital city of Santiago. The box plots highlight the exceptional growth this city has undergone, particularly since 1970. The growth coincides with a decade of tremendous changes—Pinochet's accession to power, abandonment of the Import Substitution Industrialization (ISI) model, and growth of the tertiary sector to the detriment of industry, which was progressively concentrating in Santiago (Henríquez, Azócar, & Romero, 2006). We also observe an increasing number of large cities surpassing the upper hinge of the box plot (1.5 times plus the corresponding third quartile of the distribution). These could be the first evidence of general agglomeration economies in Chile, although further analysis on structural changes and spatial regimes should be done to confirm this conclusion.

The figures show that evolution of the population has indeed not been homogeneous throughout the country or by sub-periods. As in other studies (e.g. Longhi, Musolesi, and Baumont, 2014), we thus proceed to exploratory data analysis. This analysis shows different behavior of the Metropolitan Region (MR) cities, which have moved at a faster pace than other regional

[2]The Republic of Chile is politically divided into regions, provinces, "comunas" (municipalities) and censal districts. Within each municipality, there are different "entities": cities, towns, villages, hamlets and others.

[3]Despite the existence of previous censuses, the first period of analysis is 1930 because of the foundation of the Southern city of Aysen in 1928. We decided to include it in the sample because, in the far South of Chile, cities are scarce and disseminated. We also decide to exclude information derived from the 2012 Chilean census due to some detected important methodological problems. As regards the new 2017 Census, data on entities are not still available. For more information see (Instituto Nacional de Estadistical, 2014).



(a) Location (b) Growth

FIGURE 5.1: Location and population growth of the Chilean cities
Source: self-elaboration

cities, particularly since the 1970s (Figure 5.2(a))[4]. Both when we include Santiago in the MR and when we exclude it, however, we see differences in development of its satellite cities relative to the rest of the country (Figure 5.2(b)). The metropolitan area's current cities were smaller towns for a long time, with a population size significantly below the national average, but this situation changed abruptly during the 1980s, a time of significant structural change and outstanding growth due to Santiago's sprawl. This evolution is consistent with that of many other world metropolitan cities (e.g. Black and Henderson, 2003 for US MSAs and Lanaspa, Pueyo, and Sanz, 2003 for Spanish cities), due to population spillover effects caused by negative agglomeration externalities in the capital (pollution, lack of green spaces, high housing price, etc.).

[4]All computations, produced primarily with the R package, are available upon request from the authors.

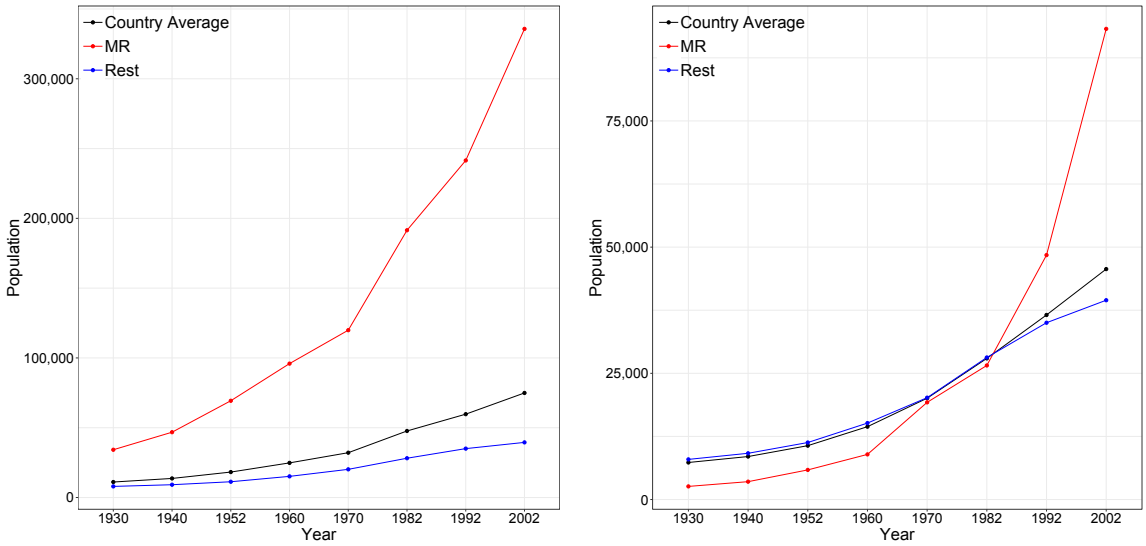
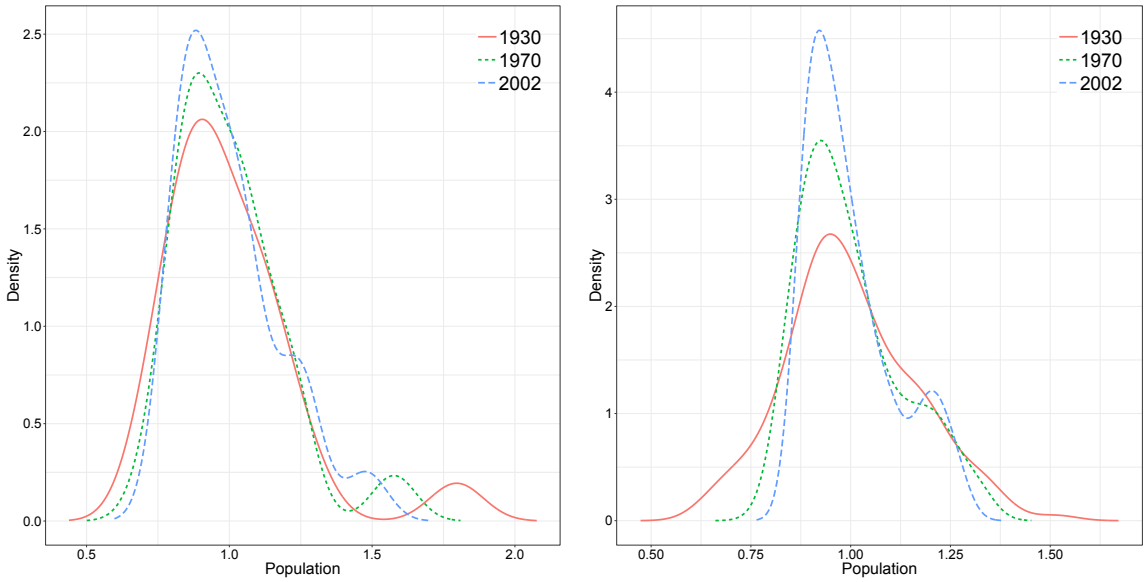


FIGURE 5.2: Population growth by structural changes and spatial regimes
Source: self-elaboration
Note: MR is Metropolitan Region, Rest is rest of the country, and the black line corresponds to the country averages.

5.2.2 The evolution of the shape of the Chilean urban population distribution

Typical spatiotemporal exploratory data analysis does not permit inferences about patterns in the intertemporal evolution of the full cross-sectional distribution of cities in terms of size relative to the rest of the urban system (Quah, 1996). We thus follow a strand of the literature that estimates non-parametric kernel density of urban population distributions for different periods (Le Gallo & Chasco, 2008). Specifically, we examine relative city size distribution at present (2002) and the way this distribution has changed since 1930 to analyze its characteristics of mono- or multimodality. We consider relative size distributions by normalizing the log of population size for each decade, divided by the log average size. Using the previously found evidence of both a structural change during the 1970s and two spatial regimes defined by the MR and the rest of Chilean cities, Figure 5.3 shows the kernel distributions in 1930, 1970, and 2002 for these two regimes (5.3(a) and 5.3(b), respectively). On the horizontal axis, the value 1 indicates average city size in Chile, 1.5 a value 50% higher than this average, and so on.

Some interesting results highlight the different evolution of the MR cities with respect to the rest of the country's urban system. First, in both urban groups (with and without the MR), the central mass of distributions increases significantly in 1970, peaking in the 2002 distribution. This progressive concentration of probability mass can be interpreted as



(a) Metropolitan Region (b) Rest of the Country

FIGURE 5.3: Densities of log relative urban municipality size in Chile
Source: self-elaboration

evidence of population convergence, consistent with the evolution of many other developed and developing countries (e.g. Anderson and Ge, 2005 in China or Nitsch, 2001 in some European countries). On the one hand, the convergence trend is more acute in the non-MR city group (Figure 5.3(b)), although not uniformly so, due to an observed second mode starting in 1970 that corresponds to a club of regional cities converging to a higher population mean. This outcome registers diversification of urban systems (Nitsch, 2001) due to increase in the size of mid-size cities; in fact, some mid-size Chilean cities have grown at greater rates than large cities (Henríquez et al., 2006). On the other hand, the initial club of large cities present in the MR at the beginning of the period decreases gradually, almost disappearing in 2002 (Figure 5.3(a)). The results thus give clear evidence of peri-urban growth of the MR due to steady expansion of constructed area around the urban core of Santiago (Puertas, Henríquez, & Meza, 2014). Additionally, the MR distribution becomes trimodal in 2002: We observe two additional groups of larger cities adjacent to the main mode, with sizes 30% and 50% over the system average, respectively. These groups demonstrate dominance of the largest city over the urban system distribution, a phenomenon also observable in some European countries (Nitsch, 2001).

5.3 The "estdaR" package

The "estdaR" package is a new R package to perfume exploratory spatio-temporal data analysis (ESTDA). Despite having an impressive amount of available spatial packages in R, none

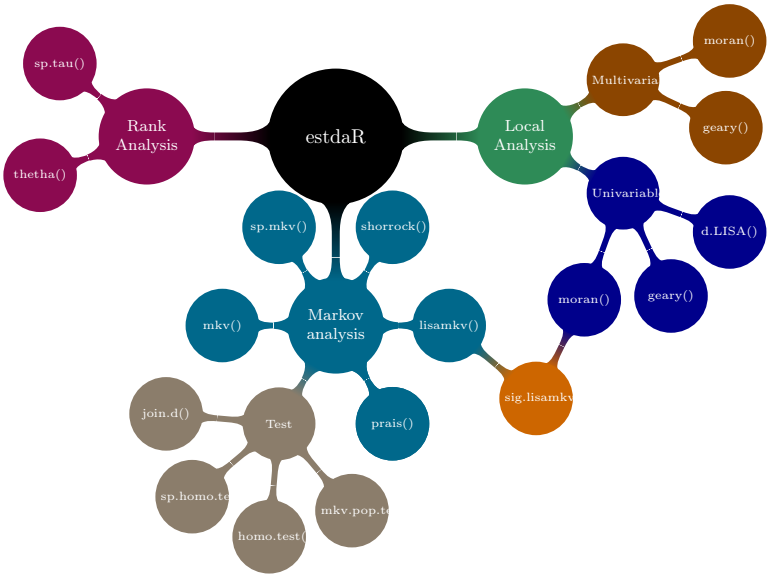


FIGURE 5.4: The 'èstdaR" packages modules
Source: self-elaboration

of them contain a set of function for the estimation of the spatial method containing in the Spatial Dynamics module of PySAL (Rey & Anselin, 2010). PySAL is an open source library for spatial analysis written in the object-oriented language Python, and is freely available to use, however, for a researcher who uses a different programming code than Python, use this library imply a significant time investment in learning a new programming language. Taking advantage of the open source feature of studying and change the source code to suit the user's needs (Stallman, 2002), we implement the PySAL's spatial dynamics module in R and we complement it with a new set of functions. The Figure5.4 shows the structure of the "estdaR" package (Vallone et al., 2018). The package contains several modules, the `mkv()` function performs the estimation of the traditional Markov probability matrix, this module include the function `sp.mkv()` to estimate the Spatial Markov probability matrix (Rey, 2001) and the `lisamkv()` to the estimation of the LISA Markov approach (Rey & Janikas, 2006). This module also contains a set of function to test the markovian properties of the discretization used by the test proposed by Bickenbach and Bode, 2003 with the `homo.test()` and the `mkv.prop.test()` functions. The homogeneity across the space of spatial Markov matrix can be testing using the test proposed by Rey, Kang, and Wolf, 2016 with the `sp.homo.test()` function. The test of the join dynamics between the spatial unit and its neighbours of the LISA Markov matrix is also possible by the use of the `join.d()` function. The function `sig.lisamakv()` which estimate a modify LISA Markov matrix (Rey & Janikas, 2006) considering a Markov chain of five classes, four classes for each significant position of the observations in the Moran's scatterplot and an extra class to capture the non-significant observations. The second module of the package contains a set of function

to perform rank method analysis. The function `theta()` estimate the Rank Decomposition index (Rey, 2004) and the `sp.tau()` function compute the Global Indicator of Mobility Association (GIMA) (Rey, 2016). The third module is centered on the local analysis. This module includes the function `d.LISA()` to estimate the Directional LISA approach (Rey et al., 2011), the functions `moran()` and `geary()` allows the user to analyze the time dimension of the traditional ESDA local Moran and local Geary c indicators (Anselin, 1995) when spatio-temporal data is used and the estimation of the Bivariate Local Moran indicator (Anselin, 1995) and the Multivariable Geary indicator (Anselin, 2017) when multivariable data is used. Following Rey, 2018 instead of considering the open source as only a research tool, we consider it as a research object as well as a pathway to better geospatial education and research, so an extra effort to keep the code the simpler as possible was made.

5.4 Mobility within the Chilean urban system between 11930 and 2002

5.4.1 Analysis of urban dynamics using Markov and Spatial Markov chains.

We first estimate the classical first-order Markov probability transition matrix, which provides information about the movements of Chilean cities within this distribution over the course of the study period. As population is a continuous variable, we must discretize the continuous state-space of this variable. Since improper discretization could have the undesired effect of removing the Markov property and producing very misleading results (Le Gallo & Chasco, 2008), we use three discretization methods based on distribution quantiles, quintiles, and deciles. Finally, we discretize this variable using quintiles because they satisfy two conditions. First, the initial classes include a similar number of observations (Le Gallo, 2004) (296 observations per class, with the exception of 3 classes with 288 observations). Second, this discretization performs best in the first-order test for Markovian property (Bickenbach & Bode, 2003). The probability maximum likelihood estimator is defined as:

$$\hat{p}_{i,j} = \frac{\sum_t n_{i,j,t}}{\sum_t \sum_j n_{i,j,t}} \tag{5.1}$$

where $n_{i,j,t}$ is the number of times a sample chain started in state i in period t and transitioned to state j in the next period (Rey, 2015). Table 5.1 shows the results of the first-order probability matrix transition of the Chilean urban system.

These findings yield three striking observations. First, as in many other social science phenomena, path dependence plays an important role in dynamics, since the main diagonal of the matrix has the highest probability value for all states. Second, the mobility of the system focuses on small and medium-size cities: on the one hand, small and small-to-medium cities have a higher propensity to move upwards in the distribution, while the medium and medium-to-large cities are more likely to move downwards. Third, the last row of Table 5.1 shows the steady-state condition (π) tending to a uniform distribution, demonstrating lack of a dominant city size in the long-run distribution.

TABLE 5.1: Markov Probaility matrix

State	Q1	Q2	Q3	Q4	Q5
Q1	0.8456	0.1274	0.0232	0.0039	0.0000
Q2	0.1390	0.6873	0.1699	0.0039	0.0000
Q3	0.0119	0.1905	0.6984	0.0992	0.0000
Q4	0.0039	0.0000	0.1004	0.8263	0.0695
Q5	0.0000	0.0000	0.0000	0.0695	0.9305
π	0.2010	0.2010	0.1957	0.2010	0.2010

Source: self-elaboration

In the second step, we explore spatial influence on Chilean urban dynamics by estimating the Spatial Markov transition matrix (Rey, 2001). This method reports the probability of a particular transition (Down, None, or Up) conditioned by the populations of the urban municipality's neighbors at the beginning of the year. As noted in Rey, 2015, the maximum likelihood estimator of transition probabilities is:

$$\hat{p}(l)_{i,j} = \frac{\sum_t n(l)_{i,j,t}}{\sum_t \sum_j n(l)_{i,j,t}} \tag{5.2}$$

where $n(l)_{i,j,t}$ is the number of times a sample chain with a spatial lag in quintile l started in state i in period t and transitioned to state j in the next period.

The conditional transition probability matrices are reported in Table 5.2. The matrices are ordered according to the value of a chain's spatial lag at the beginning of the transition period. Hence, l(1) is the chain corresponding to the spatial lags of population values located at the lowest quintile of this distribution and l(5) the chain of spatially lagged population values in the upper quintile. Computing the spatial lags requires a spatial weight matrix. We tested several definitions for this matrix and finally selected an inverse distance weights

matrix, since distance is a key factor in city growth and spatial pattern of city sizes (Ioannides & Overman, 2004).[5]

To contrast the influence of space on the transition and the homogeneity across lagged classes, we compute two statistics: Pearson's Q test and the likelihood ratio (LR) test (Rey et al., 2016). The Q statistic takes a value of $\chi^2_{52} = 70.131$ where p-value=0.048. The LR test, for 52 degrees of freedom, takes the value 73.938 where p-value=0.024. These results allow us to reject the null hypothesis of spatial homogeneity behavior in the Chilean urban system.

We might infer a predominance of path dependence, independently of lag size, since the main diagonal probability matrix always assumes the major probability value. Different spatial behavior is ultimately detected, however, since the probability of an upward or downward move differs depending on urban area context. For example, the probability of an urban municipality moving up in the hierarchy is higher when its spatial lag contains a higher average population. In addition, the l(1) Markov chain corresponding to the smallest spatial neighbors has an absorbing state in the large city state. That is, distributional movements in the group of large cities that are surrounded by smaller towns are practically null. Finally, the probability of ceasing to be a small city increases with neighbors' size but is otherwise very likely to remain in the same class. These outcomes confirms the existence of general urban agglomeration economies in Chile, as recently observed by Soto and Paredes, 2016, in line with Duranton's hypothesis (Duranton, 2016) for less-developed countries.

The steady-state distributions for neighboring cities of small and small-to-medium size, l(1) to l(3), are left-skewed distributions. This means that having small-to-medium-size cities nearby increases the probability that a city will shrink and become a smaller town in the long run. The l(4) steady-state distribution corresponding to the medium-large neighboring cities tends to a uniform distribution, as in the non-spatial case, but the l(5) chain of the largest neighboring cities has a right-skewed distribution, implying a long-term concentration process. In effect, the highest probability for a city to become a metropolis occurs when it is neighbored by very large cities.

Table 5.2: Spatial Markov transition matrix of the Chilean city population, period 1930-2002

	State	Q1	Q2	Q3	Q4	Q5
	Q1	0.7885	0.1731	0.0385	0.0000	0.0000
	Q2	0.1778	0.6000	0.2222	0.0000	0.0000
l(1)	Q3	0.0000	0.2157	0.6667	0.1176	0.0000

[5] As a robust check, similar results were obtained with other spatial weight specifications, such as driving distance and other neighborhood measures. Complete computations are available from the authors upon request.

TABLE 5.2: Spatial Markov transition matrix of the Chilean city population, period 1930-2002

	State	Q1	Q2	Q3	Q4	Q5
	Q4	0.0000	0.0000	0.0857	0.9143	0.0000
	Q5	0.0000	0.0000	0.0000	0.0000	1.0000
	π	0.1811	0.2155	0.2543	0.3491	0.0000
	Q1	0.8261	0.1304	0.0435	0.0000	0.0000
	Q2	0.1343	0.7313	0.1343	0.0000	0.0000
l(2)	Q3	0.0000	0.1905	0.7460	0.0635	0.0000
	Q4	0.0000	0.0000	0.1299	0.8571	0.0130
	Q5	0.0000	0.0000	0.0000	0.1379	0.8621
	π	0.2471	0.3200	0.2820	0.1379	0.0130
	Q1	0.8438	0.1250	0.0000	0.0312	0.0000
	Q2	0.1957	0.6957	0.1087	0.0000	0.0000
l(3)	Q3	0.0159	0.1746	0.7460	0.0635	0.0000
	Q4	0.0000	0.0000	0.1364	0.8182	0.0455
	Q5	0.0000	0.0000	0.0000	0.1045	0.8955
	π	0.3260	0.2447	0.1931	0.1646	0.0716
	Q1	0.8615	0.1231	0.0154	0.0000	0.0000
	Q2	0.1186	0.7288	0.1525	0.0000	0.0000
l(4)	Q3	0.0227	0.1591	0.7045	0.1136	0.0000
	Q4	0.0000	0.0000	0.1163	0.7209	0.1628
	Q5	0.0000	0.0000	0.0000	0.1250	0.8750
	π	0.2012	0.1995	0.1844	0.1802	0.2347
	Q1	0.8736	0.1034	0.0230	0.0000	0.0000
	Q2	0.0714	0.6429	0.2619	0.0238	0.0000
l(5)	Q3	0.0323	0.2258	0.5484	0.1935	0.0000
	Q4	0.0167	0.0000	0.0333	0.8167	0.1333
	Q5	0.0000	0.0000	0.0000	0.0256	0.9744
	π	0.0465	0.0342	0.0327	0.1430	0.7436

Source: self-elaboration

5.4.2 LISA methods

Although the SMC provides insight into the role of spatial neighbor cities at the beginning of the transition, urban system dynamics produce changes not only in cities themselves but

also in their neighboring cities. To analyze this issue, we use two different methods: the LISA Markov Chain (Rey & Janikas, 2006) and the Directional LISA (Rey et al., 2011).

LISA Markov Chain

The LISA transition matrix (Rey & Janikas, 2006) is based on the Local Moran statistic proposed by Anselin, 1995 to identify local clusters and spatial outliers. The LISA Markov chain computes the joint transition of a city and its neighbors in the distribution by measuring their movements across the four quadrants of the Moran scatterplot. Conventionally, the upper-right quadrant and the lower-left quadrant correspond to positive spatial autocorrelation (similar values at neighboring locations) and are referred to respectively as High-High (HH) and Low-Low (LL) spatial autocorrelation. The lower-right and upper-left quadrants, in contrast, correspond to negative spatial autocorrelation (dissimilar values at neighboring locations), referred to respectively as High-Low (HL) and Low-High (LH) spatial autocorrelation.

The states of the LISA Markov chains are the four quadrants of the Moran scatterplot in a given period. In each period, a city can be classified into four mutually exclusive categories HH, LH, LL, and HL where, for example, HL indicates a city above the system average for that period while its neighbors' mean size is below the average. From period to period, a city's position in the Moran scatterplot may change among the quadrants, with 16 possible transitions. Table 5.3 presents the probabilities estimated for these transitions of the LISA Markov chain corresponding to the Chilean urban system. For example, the 0.8382 value corresponding to the transition HH to HH (first cell) means that the probability that a large city with large neighbors (HH) will remain in this state is 83.82%.

A formal test for co-movement dependence –based on (Rey, Mack, & Koschinsky, 2012)– can also be performed by decomposing the LISA Markov chain into a pair of chains, one for the city and the other for the neighbors. Each chain has two states: H and L. The statistics follow an χ^2 distribution, where the null hypothesis is the independence of the two chains (Rey, 2015). In the case of Chilean cities, the statistic assumes the value χ^2_9= 31225.42 with p<0.001, allowing us to reject the null hypothesis and demonstrating co-dependence of the movement of a city and its neighbor cities.

As in the previous cases, movements are more frequent within quadrants than between them, providing new evidence of high persistence in the system (Table 5.3). Apart from the main diagonal, the most frequent movements take place from states HH to LH, which may be interpreted as a suburbanization process; and from states HH to HL, evidence of population concertation processes (Sayas, 2006). The steady-state condition (π) shows higher probability values for HL and LL and lower values for HL and HH, indicating that only a few

large cities (H state) and a lot of small ones (L state) will exist in the long run. The Chilean urban system thus exhibits a clear and persistent pattern of agglomeration economies. The increase in the size of the intermediate cities found in the density plot demonstrates a regional concertation process. Population concentration thus occurs not only in the MR of Santiago, but also in the main important city at regional level. This finding aligns with evidence for Latin American Countries such as Mexico (Pimentel, 2000) and Brazil (Baeninger, 1997), where regional growth trends spread beyond large metropolitan areas but usually maintain a high degree of demographic concentration in large and medium-sized cities, especially in large metropolitan areas (Da Cunha, 2003, 2013).

TABLE 5.3: LISA transition matrix of the Chilean city population, period 1930-2002

	HH	LH	LL	HL
HH	0.8382	0.1029	0.0000	0.0588
LH	0.02266	0.9589	0.0164	0.0021
LL	0.0000	0.0270	0.9696	0.0034
HL	0.0426	0.0000	0.0567	0.9007
π	0.0833	0.4612	0.3835	0.0720

Source: self-elaboration

Directional LISA

From period to period, it is possible to find, for example, that a small city with a high growth rate is surrounded by small neighboring cities that are also growing rapidly even though they are below the mean city size. That is, these cities grow, but not enough to change their position in the Moran scatterplot. The LISA Markov method considers this city as a 'static' city, in the sense that it does not move across the Moran scatterplot states. One way of capturing the co-movements of cities and neighbors across the Moran scatterplot is the Directional LISA approach (Rey et al., 2011). This method visualizes these co-movements by means of the origin-standardized movement vector (Rey et al., 2011), obtained by comparing two Moran scatterplots corresponding to two different periods of time. Figure 5.5 represents the standardized Directional LISA of the Chilean cities, in which the movement vectors reflect relative changes in the LISA Markov chain between the first and last period of analysis (1930 and 2002, respectively). We have standardized the moves such that all arrows depart from the coordinate origin of the Moran scatterplot. For example, movements to the 'Southwest' part of the scatterplot indicate a reduction in a city's size concurrent with reduction in its

neighbors' size. Similarly, movements to the Northeast represent an increase in the size of both the city and its neighbors during the period.

This technique is very appropriate to test for different dynamics between the MR and the rest of the country, as Gregory and Patuelli, 2015 do for the German regions. In the previous Kernel distribution analysis, we find some evidence of two different spatial regimes with the MR and the other regional Chilean cities—exactly what we find in Figure 5.5. Red arrows indicate MR dynamics and blue arrows cities in the rest of the country. While behavior of the regional city's growth during this long period is diverse, with movements at any direction, the cluster of MR cities exhibits a striking movement pattern towards the 'North' and 'Northeastern' parts of the Moran scatterplot. Nearly all MR cities experienced concurrent population growth with their neighbors. Only a few cities diverged from their neighbors, with size reduction and increase, respectively. This last result is consistent with the outcome of the previous exploratory analysis of vast urban sprawl from Santiago to its satellites, which occurred primarily during the 1980s and corresponds to the so-called 'concentrated deconcentration' phenomenon (Rodriguez, 2007).

To obtain a clearer view of the movement patterns, we build two rose diagrams (Gutiérrez & Rey, 2013; Rey et al., 2011) (See Figure 5.6). The rose plot, based on circular statistics, is a circular histogram that shows the frequency of moves across different directions based on the angular notation. We produce two rose diagrams for the Chilean cities, divided into 8 classes: one for 1930-1960 (Figure 5.6(a)) and the other for 1970-2002 (Figure reff4-5b).

Analysis of the rose diagram shows different spatial and temporal behavior of the MR regime. The most frequent movement in the RM regime is the 0-to-45 degree direction, showing major growth in the different cities with respect to their neighbors—a clear process of population concentration in this regime. The rest of the country's most significant movement takes place in the 180-to-225 degree direction, meaning that regional cities experienced greater reduction in size than their neighbors, indicating the presence of intra-urban migration. Analyzing the overall systems reveals an inter-urban migration process; all movement 90-270 implies a reduction in the city size and a repulsion movement from these cities to another city in the system. In the Chilean urban system, as in other Latin American countries, spatial mobility over shorter distances impacts the demographic growth of many cities, including medium-sized towns, especially municipalities located in the outlying areas of large urban centers (Da Cunha, 2013; Rodríguez & Da Cunha, 2009).

In the period 1930-1960, depicted in Figure 5.6(a), the MR regime shows a small group of cities in the 270-360 degree class growing at the expense of their neighbors, which decrease in size. Note also a small portion (less than the 10%) of the movements in the 180-270 degree class, corresponding to a set of cities shrinking at the same time as their neighbors. In the 1970-2002 period (Figure 5.6(b)), however, all moments of the MR regime are in the 0-180

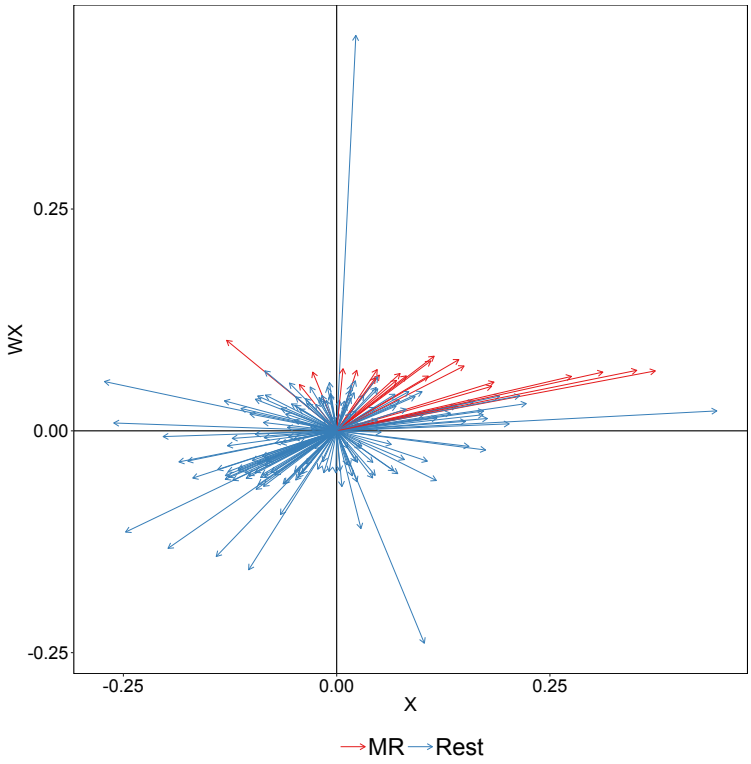
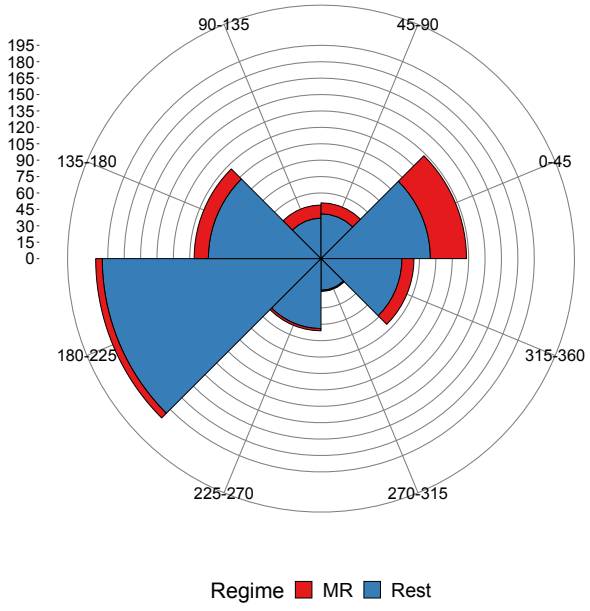


FIGURE 5.5: Standardized directional Moran Scatterplot of the Chilean cities by spatial regimes between 1930 and 2002
Source: self-elaboration

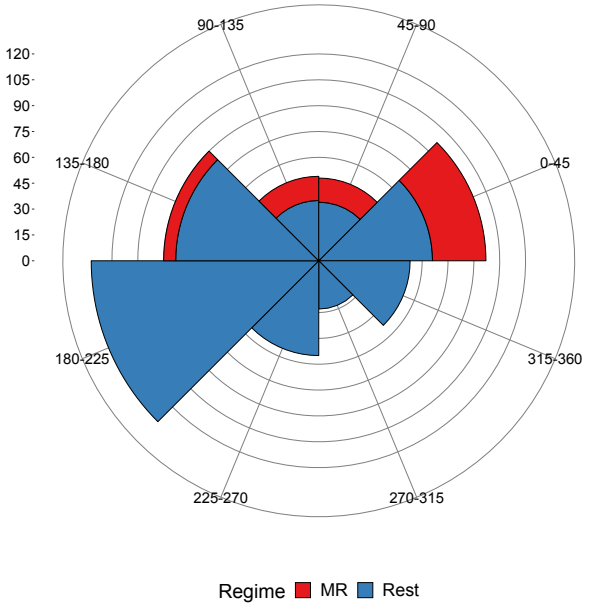
degree class. This region thus includes a spatial sub-cluster of neighboring cities growing at the same pace (0-90 degree class) and coexisting with other city sub-clusters—that is, undergoing suburbanization (90-180 degree class). These movements are consistent with peri-urban growth of Santiago City and demonstrate the presence of a strong agglomeration process in the MR region. The cluster of regional cities follows a similar movement distribution in both periods. The most frequent direction movement takes place in the 180-270 degree class (the 'Southeast' direction), followed by the 0-90 degree direction ('Northwest' direction). Hence, the most common directional movement in the regional cluster is repulsion, characterized by joint reduction in the size of both a city and its corresponding neighbors. We also observe a movement of less intensive absorption, in which a concurrent city grows at the expense of its surrounding areas.

5.4.3 Global Indicators of Mobility Association (GIMA)

To continue our exploration of different spatial regime behaviors in the Chilean urban system, we use the Global Index of Mobility Association (GIMA) (Rey, 2016), whose antecedent is the Kendall spatial τ presented by Rey, 2004. This indicator is based on the Kendall rank



(a) 1930-1960



(b) 1970-2002

FIGURE 5.6: Rose diagram
Source: self-elaboration

correlation coefficient (Kendall, 1962), which measures concordance of the ranks of a variable in two periods of time $\tau\left(y_t,\ y_{t+1}\right)$ as follows:

$$\tau\left(y_t,\ y_{t+1}\right) = \frac{c-d}{\frac{n\left(n-1\right)}{2}} \tag{5.3}$$

where c is the number of concordant pairs, d the number of discordant pairs, and n the sample size. The τ index ranges from 0 (perfect discordance) to 1 (perfect concordance). From this expression, we can build a mobility index as follows:

$$M = \frac{\tau\left(y_t,\ y_{t+1}\right) - 1}{-2} \tag{5.4}$$

M varies from 0 to 1, where 1 implies full ranking mobility and 0 complete stability of the ranking.

It is thus possible to decompose the τ index to capture the effect of space in the ranking changes as follows[6]:

$$\tau\left(y_t,\ y_{t+1}\right) = \varphi\tau_W\left(y_t, y_{t+1}\right) + \left(1-\varphi\right)\tau_{_W}(y_t, y_{t+1}) \tag{5.5}$$

where $\varphi = \frac{i'Wi}{i'\left(W+W\right)i}$, being i a unit vector of order (nx1), W a spatial weights matrix containing the neighboring relationships and $W = ii^{'} - W - I_{n\times n}$ a matrix capturing the non-neighboring relationships.

Equation 5.5 presents the spatial τ index of concordant and discordant rank pairs as a decomposition of two τ indexes: one for pairs of neighboring observations and the other for pairs of non-neighboring observations. This procedure enables us to identify the different correlation patterns between neighboring and non-neighboring cities. Different ranking patterns may be inferred based on random spatial permutations of the attributes to develop a distribution for τ_W under the null hypothesis of spatial homogeneity in the correlation patterns (Rey, 2016).

As with the Kendall's τ, it is possible to construct a spatial mobility index as follows:

$$M_W = \frac{\tau_W\left(y_t,\ y_{t+1}\right) - 1}{-2} \tag{5.6}$$

[6] Rey, 2016 presents the full mathematical decomposition of this index.

Equation 5.6 also allows for additive decomposition of overall mobility, which gives the option of comparing different levels of mobility between neighboring and non-neighboring cities, as follows:

$$M = \varphi M_W + (1 - \varphi) M_{_W} \tag{5.7}$$

In Table 5.4, we report the mobility and τ index decomposition, with their corresponding p-values[7], for the Chilean urban system to detect differences in behavior between the RM and the regional cities, using a regime weight matrix in the computation.

TABLE 5.4: Spatial Kendall indexes for the Chilean city population in the period 1930-2002, by spatial regimes

Period	M_W [8]	M_{noW}	τ_W	τ_{noW}	p-value
1930-1940	0.0781	0.0678	0.8438	0.8644	0.2920
1940-1952	0.0805	0.0764	0.8390	0.8472	0.3990
1952-1960	0.0651	0.0721	0.8698	0.8558	0.2460
1960-1970	**0.0625**	**0.0920**	**0.8750**	**0.8159**	**0.0160**
1970-1982	**0.0492**	**0.0680**	**0.9017**	**0.8640**	**0.0160**
1982-1992	0.0540	0.0548	0.8920	0.8904	0.4560
1992-2002	0.0551	0.0567	0.8898	0.8866	0.4350

Source: self-elaboration

For only two periods, 1960-1970 and 1970-1982, are the spatial indexes statistically significant and the spatial τ_W, corresponding to the MR cities, significant. This result demonstrates differences in ranking mobility in space that coincide with structural change in the 1970s in Chile. During this period, the ranking mobility inside this regime (M_W) was smaller than the mobility outside the regime (M_{noW}), implying more concordant movement of the cities belonging to the MR regime than of those in the rest of the country. A change in the rank of the MR cities is thus more likely to be in the same direction as in the regional cities, which underwent more changes in opposite directions. The synchronic evolution of the MR cities during these 22 years is consistent with the reinforcement this area experienced due to abandonment of the ISI policy and trade liberalization. In effect, these changes in the national policy generated a strong migration trend from rural areas to intermediate and large cities with production structures based on comparative advantage goods (Escolano Utrilla

[7] In this paper, computation of the p-values is performed with a 1,000-replication process.

[8] **Note:** subscripts 'W' and 'noW' mean the MR and non-MR spatial regimes

et al., 2007; Geisse & Valdivia, 1978; Geisse, 1977), producing strong agglomeration forces around the MR cities.

5.4.4 Rank decomposition

The rank decomposition index $\Theta_{t_1-t_0}$ (Rey, 2004) is defined as the sum of rank changes, from period t_0 to t_1, within a regime over the sum of the overall rank changes. Formally, if we set $\theta_{i,t}$ as the position in the ranking of city i in period t and assume the existence of R spatial regimes, the rank decomposition index Θ is calculated as follows:

$$\Theta_{t_1-t_0} = \frac{\sum_{R} \left| \sum_{i \in R} \theta_{i,t_1} - \theta_{i,t_0} \right|}{\sum_{i} \left| \theta_{i,t_1} - \theta_{i,t_0} \right|} \tag{5.8}$$

The cohesion index will take the value 0 in the case of complete absence of cohesion (i.e., when all changes in the ranking occur only inside the same regime). At the other end, Θ will take the value 1 when all movements in the ranking are 100% 'cohesive' within the regimes. In this context, cohesion can be understood as a process of migration between regimes, such that the size of cities in one regime increases (ascending in the ranking), at the expense of the size of cities in the other regime, which decline in the ranking. Full cohesion thus implies a perfect population transfer between regimes.

Although Θ is a non-parametric test, it is possible to construct an inferential process based on random spatial permutations under the null hypothesis of spatial homogeneity (Rey, 2004). Table 5.5 reports decomposition and its p-value. As in the previous method, we used the two Chilean urban regimes defined by the MR and the rest of the regional cities.

TABLE 5.5: Rank decomposition index Θ for the Chilean urban population in the period 1930-2002, by spatial regimes

Period	Θ	p-value
1930-1940	0.0211	0.7900
1940-1952	0.1191	0.1060
1952-1960	**0.1548**	**0.0430**
1960-1970	**0.2912**	**0.0000**
1970-1982	0.0794	0.2410
1982-1992	**0.1396**	**0.0460**
1992-2002	**0.1994**	**0.0040**

Source: self-elaboration

During the periods 1952-1970 and 1982-2002, the Θ index is statistically significant and takes the highest values, implying a stronger migration process between regimes (Table 5). One drawback of this index, however, is that it cannot provide information about the direction of the migration flow. As stated before, specific events in Chile's history suggest that the MR of Santiago was the main beneficiary of this transfer. In the period 1952-1970, agricultural reform and the ISI policy produced rural-urban migration in Chile that benefited the MR growth (De Mattos, 1999; Geisse & Valdivia, 1978; Geisse, 1977). During 1982-2002, the MR of Santiago had a selective agglomeration process based on higher income and better amenities, which attracted migrants with high human capital (Cambiaso, Alonso, & Claro, 2001).

5.5 Conclusions

In this paper, we use a set of novel tools to improve knowledge of the Chilean urban system as a whole, overcoming certain limitations in the existing literature on Latin American countries, which tends to study the capital city and its metropolitan area and to disregard the rest of the regional urban system. We firmly conclude that real overconcentration of population and economic activities in these metro areas should not obscure investigation of the regional cities, which they have their own idiosyncrasies and do not necessarily follow the same spatiotemporal trends.

We employ some novel statistical techniques not previously applied to urban studies and revealed to be very useful for detecting spatial differences in the evolution of an urban system. Intelligent combination of these tools, which can be computed with 'estdaR' (an R package), answers some questions posed in the introduction of this paper.

First, are certain urban processes, such as urban sprawl and population convergence, homogeneous across Chile's entire city system? The answer is no. Initial exploratory data analysis shows a double structural change due to the existence, first, of two spatial regime clusters defined by the Metropolitan Region (MR) and Chile's other cities; and, second, two sub-periods defined by the 1970s. The MR city area exhibits different behavior—growing faster than the regional cities due to the existence of vigorous urban sprawl from Santiago city to its satellites. Estimating the kernel distributions shows a peri-urban process in the MR (as in many developed countries), which took place during the 1970s and 1980s, coinciding with tremendous economic and political changes. This process is creating population convergence inside this spatial regime, although Santiago clearly continues to predominate. The regime formed by the other regional cities is also undergoing population convergence, but at a higher speed than the metro area.

Second: Is Duranton's hypothesis (Duranton, 2016) about the existence of stronger agglomeration effects in less-advanced countries applicable to Chile? The answer is yes. On the one hand, estimating the standard and SMC matrices highlights the strong role played by path dependence in the Chilean urban system: large cities tend to remain large and small ones to remain small or to grow. Additionally, the LISA Markov method demonstrates the existence of processes of both suburbanization (e.g. in the MR area) and concentration in the regional city group. In the long run, Chile will have only a few large cities and many small ones, as its urban system shows a clear and persistent pattern of agglomeration economies.

Third, will a city population grow faster or slower depending on its neighbors' growth speed? The estimation of the spatial Markov and LISA Markov matrices enables us to conclude that the probability a city will grow increases with its neighbors' size, while large cities surrounded by smaller towns hardly experience any change in population. Spatial proximity thus matters in the urban system, usually by promoting agglomeration.

In its urban policy, Chile must recognize two different city clubs: the MR cluster, dominated by the city of Santiago; and the other regional cities, which share their own characteristics and dynamics. The MR cities are converging to Santiago at a slower speed than the regional cities are converging among themselves. In the long run, therefore, Chile will have only a few large cities and a lot of small ones, demonstrating the national urban system's clear and persistent pattern of agglomeration economies. A more balanced regional urban system, growing at the expense of the nearby rural areas, will coexist with a highly concentrated MR, in which Santiago's population spills over into its satellites, moving towards (still distant) future convergence.

6 Conclusions

Geocomputation can be defined as "[t]he art and science of solving complex spatial problems with computers" (Cheng et al., 2012). The increase in large-scale computation caused by this technological advance has increased the importance of the term geocomputation (Batty, 2017). Since geocomputation's focus is on spatial analysis, the spatial economic and regional sciences quickly adopted geocomputational techniques to study the complex structures of urban and regional systems. This dissertation aims to contribute to the use of geocomputation in spatial economic analysis through the construction and application of a new set of algorithms and functions in the R programming language to analyse spatial economic data.

Chapter 3 presents the "DataSpa" R package. "DataSpa" is a new set of geocomputational tools to solve problems of data accessibility and availability in data collection. Accessibility problems occur when the way data are published in servers blocks or delays the download process, frequently causing errors in the construction of large databases. Availability problems usually arise when official agencies restrict access to the information, producing empty data records and incomplete databases. Overcoming these problems requires development of new data extraction strategies and exploration of new information sources.

Using the URL parsing strategy, we build a set of functions to download, load and manipulate population and unemployment databases to solve the accessibility problems presented by two official agency web portals. We solve accessibility problems in the vehicle fleet database by combining URL parsing and PDF extraction strategies. First, we build the `parquet.aut()` function, which employs a URL parsing strategy to download the PDF files of municipality reports from the DGT web portal in order to extract the statistical data through a PDF extraction strategy. Next, constructing the firm databases requires solving availability problems. We apply a web scraping strategy with the functions `data.firm()` and `data.firm.a()` to download company and freelancer information freely published by a private company. Creation of a firm database provides very helpful knowledge of the distribution of economic activities in Spain at municipal and individual level.

This package has some limitations. Although the methodology and strategy used are generalizable, the code itself is not. The web scraping and pdf extraction techniques are case-adapted, and each data collection process requires a unique computational code. In addition,

data collection is strongly dependent on the stability of the source website's URL structure. If one package information source changes its domain or its storage infrastructure, the download functions will collapse, requiring new function coding to ensure package functionality

The information collected by the "DataSpa" package makes it possible to pursue a large set of economics research topics, for example, to study clusterization of municipal spatial economic activities using the dataset of firms or to study the evolution and characteristics of a vehicle fleet at municipality level. This data extraction technique also makes it possible to collect a new set of information, such as housing prices and housing characteristics or used car prices and related characteristics, to support or conduct new spatial economics research.

Chapter 4 presents another geocomputational tool to solve data accuracy problems. We use a search and replace algorithm to build the `msp()` function that interactively harmonizes data with problems such as spelling errors, acronym abbreviations and differently listed names. This function enables us to harmonize a novel dataset of more than 42,000 records collated from the Chilean National Institute of Industrial Property to study patenting activity and research collaboration in Chile between 1989 and 2013. First, we perform a full descriptive analysis of the Chilean patent ecosystem. Second, we use social network analysis to understand the structure and dynamics of collaborative research production in the country. We also adapt classical spatial econometrics techniques to network visualization to explore the existence of productivity spillovers between the network's members and to contrast the relationship between productivity and collaboration using the adjacency matrix as 'vicinity' weights.

Evidence reveals intensification of knowledge content out of domestic and foreign companies reflected in a growing number of patents granted in the country but also a little small contribution of universities to the total percentage of such patents. In addition, the study results reveal long and variable patent-grant lags, as well as lack of collaboration among industrial companies and between science and industry. The clear preponderance of companies and universities without any ties at all in Chile shows the country's patent ecosystem to be poor and highly concentrated in groups, mainly around large non-resident companies. Results for the productivity-collaboration relationship show the absence of productivity clusters and null productivity spillovers in the small collaborative assignee subset.

We recognize that this research has some limitations. First, since the algorithm may find two similar valid entries in the database, we designed the `msp()` function to be interactive. This design may be a source of errors. If the data set is very large, replacement selection becomes repetitive and may generate mistakes in the harmonization process. Second, we worked with a truncated dataset. The study is limited to patents valid to 2014 and thus limits the presence of old data. Additionally, since we use a small subset of the total number

of patents filed in Chile, the database only includes patents for inventions that fulfil the criteria of novelty, usefulness and non-obviousness.

This study analyses the company's productivity and collaboration, but the dataset contains information on the patent's inventors. Future research will use this inventor's information to measure the company's productivity and collaboration pattern. The research also raises a set of interesting questions for future study, such as the existence of hub inventors who function as a nexus between companies and universities, and the presence of local spillovers in the inventor's collaborative network.

Chapter 5 presents another set of geocomputational tools to perform spatio-temporal analysis. The package "estdaR" enables analysis of the spatio-temporal dynamics of a variable, as well as clustering and detection of the spatial cluster dynamic of a variable. We use these tools to analyse the spatio-temporal dynamics of the Chilean urban system between 1930 and 2002. First, we characterize the cross-sectional distribution of existing cities using non-parametric estimations of density functions for a set of significant years. Second, we model the growth process as a first-order stationary Markov chain and evaluate the effect of spatial autocorrelation on the transition probabilities using a set of indices based on the spatial version of the standard Markov chain. Third, we perform in-depth analysis to detect spatial regimes in the movement direction and ranking mobility of Chile's urban distribution.

The results provide evidence of spatial heterogeneity in Chile's urban system, particularly stronger agglomeration effects within this system. Analysing the Spatial Markov matrices highlights the prominent role of path dependence in Chile's urban system. The ergodic distribution of the LISA Markov matrix shows a clear, persistent pattern of agglomeration economies. We also find two different regimes in Chile's urban system, defined by the Metropolitan Region (MR) and Chile's other cities. Analysis of the GIMA indicator and Rank decomposition reveals different spatial dynamics of these regimes in different time periods.

Although the methods used enable us to detect different regional behaviours in Chile's urban system, they have some limitations. The ranking decomposition method shows the presence of some population migration between the regimens used, but we cannot determinate the direction of the flow. A similar limitation occurs in the Directional LISA method. We can detect movement between 90 and 180 degrees, indicating that a city's size decreases while its neighbours' size increases, a sign of suburbanization. It is not possible, however, to confirm that this movement is a pure suburbanization process because the method does not permit us to determine the source of the neighbours' growth. City i's neighbours may increase the size of city i due to attrition from city i's expulsion flow (a pure suburbanization process) or absortion of a non-neighbour city j's migration flow. A similar condition occurs with the

urbanization process. In both cases, only history provides additional information to answer our research questions.

The package "estdaR" opens new research possibilities, as it enables the application in R of a new set of tools for spatio-temporal analysis. This Dissertation applies only a small number of the package tools, and the package contains a module that treats the local multivariable analysis and local temporal analysis not used in this study. Further, the package itself is new research. We plan to publish a paper that presents the principal characteristics of the packages and demonstrates their functionality.

7 Conclusiones

El geocomputation se puede definir como "él arte y ciencia de resolver problemas espaciales complejos con las computadoras" (Cheng et al., 2012). El incremento en la capacidad de cómputos a gran escala causado por el avance tecnológico incrementa la importancia del término Geocomputación (Batty, 2017). Dado que la Geocomputación se centra en el análisis espacial, la economìa espacial y la ciencia regional rápidamente adoptaron sus técnicas para estudiar las complejas estructuras de los sistemas urbanos y regionales. Esta tesis doctoral pretende contribuir al uso de la Geocomputación en el análisis económico espacial a través de la construcción y aplicación de un nuevo conjunto de algoritmos y funciones en el lenguaje de programación R para analizar datos económicos espaciales.

El capitulo 3 presenta el paquete de R “DataSpa”. “DataSpa” es un nuevo conjunto de herramientas geocomputacionales para resolver problemas de accesibilidad y disponibilidad de datos en la recolección de datos. Los problemas de accesibilidad se producen cuando los datos se publican en servidoresde forma que bloquean o retrasa el proceso de descarga, causando con frecuencia errores en la construcción de grandes bases. Los problemas de disponibilidad suelen surgir cuando las agencias oficiales restringen el acceso a la información, produciendo registros de datos vacíos y bases incompletas. La superación de estos problemas requiere el desarrollo de nuevas estrategias de extracción de datos y la exploración de nuevas fuentes de información.

Utilizando la estrategia de URL parsing, construimos un conjunto de funciones para descargar, cargar y manipular bases de datos de población y desempleo resolviendo los problemas de accesibilidad presentados por los portales web de los organimos oficiales. Solucionamos problemas de accesibilidad en la base de datos del parque automotor combinando estrategias de URL parsing y PDF extraction. Primero, construimos el `parquet.aut()` , que emplea una estrategia de URL parsing para descargar los archivos PDF de los informes municipales desde el portal web de DGT con el fin de extraer los datos estadísticos a través de una estrategia de PDF extraction. A continuación, la construcción de bases de datos de empresas requiere resolver problemas de disponibilidad. Aplicamos una estrategia de web scraping en las funciones `data. firm ()` y `data. firm. a ()` para descargar la información de las empresas y los autónomos publicada gratuitamente por una empresa privada. La creación

de una base de datos de empresas proporciona un conocimiento muy útil de la distribución de las actividades económicas en España a nivel municipal e individual.

Este paquete tiene algunas limitaciones. Aunque la metodología y la estrategia utilizadas son generalizables, el propio código no lo es. Las técnicas de PDF extraction y web scraping se adaptan a cada caso, por tanto cada proceso de recopilación de datos requiere un código computacional único. Además, la recopilación de datos depende fuertemente de la estabilidad de la estructura de URL del sitio web de origen. Si una fuente de información del paquete cambia su dominio o su infraestructura de almacenamiento, las funciones de descarga se colapsarán, lo que requerirá una nueva codificación de función para garantizar la funcionalidad del paquete.

La información recolectada por el paquete "DataSpa" permite llevar a cabo un gran conjunto investigaciones económicas, por ejemplo, estudiar la clusterización de las actividades economicas a nivel municipal utilizando la base de datos de las empresas o estudiar la evolución y características del parque automotor a nivel municipial. Esta técnica de extracción de datos también permite recoger un nuevo conjunto de información afines, tales como precios y características de vivienda o precios y características automoviles usados, para complementar o conducir nuevas investigaciones de economía espacial.

El capitulo 4 presenta otra herramienta geocomputacional para resolver problemas de precisión de datos. Usamos un algoritmo de búsqueda y reemplazo para construir la función `msp()` que armoniza de manera interactiva datos con problemas como errores ortográficos, abreviaturas de acrónimos y nombres de listados diferente. Esta función nos permite armonizar un nuevo conjunto de datos de más de 42.000 registros recopilados del Instituto Nacional de propiedad industrial de Chile para estudiar la actividad de patentamiento y la colaboración en investigación en Chile entre 1989 y 2013. En primer lugar, realizamos un análisis descriptivo completo del ecosistema de patentes chileno. En segundo lugar, usamos análisis de redes sociales para entender la estructura y dinámica de la producción colaborativa de la investigación en el país. También adaptamos las técnicas clásicas de Econometría espacial a la visualización de redes para explorar la existencia de los efectos de contagio de productividad entre los miembros de la red y para contrastar la relación entre productividad y colaboración utilizando la matriz de adyacencia como matriz de pesos espaciales para medir la "vecindad".

La evidencia revela la intensificación del contenido del conocimiento en empresas nacionales y extranjeras reflejado en un número creciente de patentes concedidas en el país, pero también una pequeña contribución de las universidades al porcentaje total de dichas patentes. Además, los resultados del estudio revelan retrasos largos y variables en la concesión de patentes, así como la falta de colaboración entre las empresas industriales y entre la ciencia y la industria. La clara preponderancia de empresas y universidades sin ningún tipo

de vínculo en Chile muestra que el ecosistema de patentes del país es pobre y está altamente concentrado en grupos, principalmente en torno a grandes empresas no residentes. Los resultados de la relación productividad-colaboración muestran la ausencia de clústeres de productividad y derrames de productividad nulos en el pequeño subconjunto de empresas colaborativas.

Reconocemos que esta investigación tiene algunas limitaciones. En primer lugar, dado que el algoritmo puede encontrar dos entradas válidas similares en la base de datos, hemos diseñado la función `msp()` de forma interactiva. Este diseño puede ser una fuente de errores. Si el conjunto de datos es muy grande, la selección del registro de reemplazo se vuelve repetitiva y puede generar errores en el proceso de armonización. En segundo lugar, trabajamos con un conjunto de datos truncado. El estudio se limita a las patentes válidas para 2014 y, por lo tanto, limita la presencia de datos antiguos. Además, usamos un pequeño subconjunto del número total de patentes archivadas en Chile, la base de datos sólo incluye patentes para invenciones que cumplen con los criterios de novedad, utilidad y no-obviedad.

Este estudio analiza la productividad y la colaboración de las empresas, pero la base de datos contiene información sobre los inventores de la patente. Futuras investigaciones usará la información de los inventores para medir la productividad y el patrón de colaboración. Esta investigación también plantea un conjunto de preguntas interesantes para el futuro estudio, como la existencia de inventores que funcionan como nexo entre empresas y universidades, y la presencia de derrames locales en la red colaborativa de inventores.

El capítulo 5 presenta otro conjunto de herramientas geocomputacionales para realizar análisis espaciotemporales. El paquete "estdaR" permite el análisis de la dinámica espaciotemporal de una variable, así como el agrupamiento y detección de la dinámica de Cluster espacial de una variable. Utilizamos estas herramientas para analizar la dinámica espaciotemporal del sistema urbano chileno entre 1930 y 2002. En primer lugar, caracterizamos la distribución de las ciudades existentes utilizando estimaciones no paramétricas de las funciones de densidad para un conjunto de años significativos. En segundo lugar, modelamos el proceso de crecimiento como una cadena estacionaria de Markov de primer orden y evaluamos el efecto de la autocorrelación espacial en las probabilidades de transición usando un conjunto de índices basados en la versión espacial de la cadena estándar de Markov. En tercer lugar, realizamos un análisis a fondo para detectar los regímenes espaciales en la dirección del movimiento y la movilidad de ranking de las ciudades de Chile.

Los resultados proporcionan evidencia de la existencia de heterogeneidad espacial en el sistema urbano chileno, particularmente en los fuertes efectos de aglomeración dentro del sistema. El análisis de las matrices espaciales de Markov destaca el papel prominente de la dependencia del pasado en el sistema urbano chileno. La distribución ergódica de la matriz LISA Markov muestra un patrón claro y persistente de las economías de aglomeración.

También encontramos dos regímenes diferentes en el sistema urbano de Chile, definidos por la región metropolitana (MR) y otras ciudades de Chile. El análisis del indicador GIMA y la descomposición del ranking revela diferentes dinámicas espaciales de estos regímenes en diferentes periodos de tiempo.

Aunque los métodos utilizados nos permiten detectar diferentes comportamientos regionales en el sistema urbano chileno, tienen algunas limitaciones. El método de la descomposición del ranking muestra la presencia de cierta migración de la población entre los regímenes usados, pero no podemos determinar la dirección del flujo. Una limitación similar ocurre en el método direccional de LISA. Podemos detectar el movimiento entre 90 y 180 grados, indicando que el tamaño de una ciudad disminuye mientras que el tamaño de sus vecinos aumenta, un signo de suburbanizacion. Sin embargo, no es posible, confirmar que este movimiento es un proceso de suburbanización puro porque el método no nos permite determinar la fuente del crecimiento de los vecinos. Los vecinos de la ciudad de i pueden aumentar el tamaño de la ciudad i debido a la capatación del flujo xpulsado por la ciudad i (un proceso puro de la suburbanización) o por la absorción del flujo de una ciudad no vecina j. Una condición similar ocurre con el proceso de la urbanización. En ambos casos, sólo la historia proporciona información adicional para responder a nuestras preguntas de investigación.

El paquete "estdaR" abre nuevas posibilidades de investigación, ya que permite la aplicación en R de un nuevo conjunto de herramientas para el análisis espaciotemporal. Esta disertación aplica sólo un pequeño número de las herramientas del paquete, y el paquete contiene un módulo que trata el análisis local multivariable y el análisis temporal local mo utilizado en este estudio. Además, el paquete en sí es una nueva investigación. Planeamos publicar un artículo que presente las principales características del paquete y que demuestre su funcionalidad.

A "DataSpa" source code

Algorithm A.1: a.letter function code

```
# busca y quita las ñ y espacios en blanco y los remplaza los
   espacion por n y _


a.letter<-function(x){
    if(str_detect(x," ")==TRUE){
    x<-str_replace_all(x," ","_")
    }
      if(str_detect(x,"\u00D1")==TRUE){
    x<-str_replace_all(x,"\u00D1","N")
    }
    x
}
```

Algorithm A.2: as.numeric.factor function code

```
# trasnform a vector of factor into a numerical vector


as.numeric.factor <- function(x) {as.numeric(as.character(x))}
```

Algorithm A.3: codifica function code

```
# Imputa los codigos del INE a los muncipios



codifica<-function(x,provincia){
```

```
   p<-c("02","03","04","01","33","05","06","07","08","48","
09","10","11","39","12","13","14","15","16","20","17","18","
19","21","22","23","24","25","27","28","29","30","31","32","
34","35","36","26","37","38","40","41","42","43","44","45","
46","47","49","50","51","52")
 names(p)<-c("ALBACETE","ALICANTE","ALMERIA","ARABA","
ASTURIAS","AVILA","BADAJOZ","BALEARES","BARCELONA","BIZKAIA"
,"BURGOS","CACERES","CADIZ","CANTABRIA","CASTELLON","CIUDAD
REAL","CORDOBA","A CORU\u00D1A","CUENCA","GIPUZKOA","GIRONA"
,"GRANADA","GUADALAJARA","HUELVA","HUESCA","JAEN","LEON","
LLEIDA","LUGO","MADRID","MALAGA","MURCIA","NAVARRA","OURENSE
","PALENCIA","LAS PALMAS","PONTEVEDRA", "LA RIOJA","
SALAMANCA","TENERIFE","SEGOVIA","SEVILLA","SORIA","TARRAGONA
","TERUEL","TOLEDO","VALENCIA","VALLADOLID","ZAMORA","
ZARAGOZA","CEUTA","MELILLA")
 buscar<-subset(mun,mun[,1]==p[provincia])
 x<-simpleCap(x)
 e<-agrep(x,buscar[,4])
 salida<-buscar[e,]
 if (length(e)>1){
   s<-subset(buscar[e,],buscar[e,4]==x)
   if (dim(s)[1]==0){
     salida<-salida[which(duplicated(salida[,2])==TRUE),]
   } else{
     salida<-s
   }
 }
 if (dim(salida)[1]==0){
   x1<-unlist(strsplit(x, " "))
   d<-sapply(x1,nchar)
   e1<-agrep(x1[which(d==max(d))],buscar[,4])
   salida<-buscar[e1,]
   if (length(e1)>1){salida<-salida[which(duplicated(salida
[,2])==TRUE),]}
   }
 cod<-paste(salida[,1],salida[,3],sep="")
 if(length(cod)>1){
   cod<-cod[1]
```

```
    }
    if(length(cod)==0){
      warning("there are same municipalities without cod")
      cod<-"No macth found"
      }
    return(cod)
  }
```

ALGORITHM A.4: data.firm.a function code

```
#' @name data.firm.a
#' @rdname data.firm.a
#'
#' @title Collects information of self-employed at a
   municipality level
#'
#' @description \code{data.firm.a} generates a data frame of
   self-employment information of a particular municipality
#'
#' @param provincia one of the 52 Spanish provinces. See \link{
   getbase.pob} for details.
#'
#' @details It is an interactive function, which requires the
   selection of a particular municipality.
#'
#' @return A data frame containing the following variables for
   each self-employment: location (province, municipality,
   address), company characteristics (name, birth, legal form,
   social object), economic activity codes and self-employment
   URL
#'
#' @family firm functions
#' @examples
#' \dontrun{data.firm.a("Ceuta")}
#'
#' @export

data.firm.a<-function(provincia){
  prov<-toupper(provincia)
```

```
b <- "https://autonomos.axesor.es/informe-de-autonomo/
 provincias/"
p<-c("Albacete","Alicante","Almeria","Alava","Asturias","
 Avila","Badajoz","Baleares","Barcelona", "Vizcaya","Burgos",
 "Caceres","Cadiz","Cantabria","Castellon","Ciudad-Real","
 Cordoba","La-Coruna","Cuenca","Guipuzcoa","Girona","Granada"
 ,"Guadalajara","Huelva","Huesca","Jaen","Leon","Lleida","
 Lugo","Madrid","Malaga","Murcia","Navarra","Orense","
 Palencia","Las-Palmas","Pontevedra","La-Rioja","Salamanca","
 Santa-Cruz-De-Tenerife","Segovia","Sevilla","Soria","
 Tarragona","Teruel","Toledo","Valencia","Valladolid","Zamora
 ","Zaragoza","Ceuta","Melilla")
names(p)<-c("ALBACETE","ALICANTE","ALMERIA","ARABA","ASTURIAS
 ","AVILA","BADAJOZ","BALEARES","BARCELONA","BIZKAIA","BURGOS
 ","CACERES","CADIZ","CANTABRIA","CASTELLON","CIUDAD REAL","
 CORDOBA","A CORU\u00D1A","CUENCA","GIPUZKOA","GIRONA","
 GRANADA","GUADALAJARA","HUELVA","HUESCA","JAEN","LEON","
 LLEIDA","LUGO","MADRID","MALAGA","MURCIA","NAVARRA","OURENSE
 ","PALENCIA","LAS PALMAS","PONTEVEDRA", "LA RIOJA","
 SALAMANCA","TENERIFE","SEGOVIA","SEVILLA","SORIA","TARRAGONA
 ","TERUEL","TOLEDO","VALENCIA","VALLADOLID","ZAMORA","
 ZARAGOZA","CEUTA","MELILLA")
url<-paste(b,p[prov],sep="")
mun<-municipio.a(url)
cual<-nn.municipio(url)
display<-c(cual, paste("[",length(cual)+1,"] Todos",sep=""))
cat(display, fill=FALSE)
resp<-readline("Indique el numero del muncipio de su interes:
  ")
if ((length(cual)+1)==as.numeric(resp)){
  resp <- seq_along(cual)
} else {
  resp <- as.integer(unlist(strsplit(resp," ")))
}
set<-mun[resp]
set
lista<-lista.empresa.a(set)
cat("se analizan",length(lista),"casos \n")
```

```
  pp<-pbapply::pblapply(lista,empresa.a)
  salida <- do.call(rbind.data.frame,pp)
  return(salida)
}
```

ALGORITHM A.5: data.firm function code

```
#' @import pbapply
#' @name data.firm
#' @rdname data.firm
#'
#' @title Collects information of firms at a municipality level
#'
#' @description Generate a data frame of firm information of a
   particular municipality
#'
#' @param provincia one of the 52 Spanish provinces. See \link{
   getbase.pob} for details.
#'
#' @details It is an interactive function, which requires the
   selection of a particular municipality.
#'
#' @return A data frame containing the following variables for
   each company: location (province, municipality, address,
   geographic coordinates), company characteristics (name,
   birth, legal form, social object), main figures (number of
   employees, social capital, sales), economic activity codes
   and firm URL
#'
#' @family firm functions
#' @examples
#' \dontrun{data.firm("Ceuta")}
#'
#' @export

data.firm<-function(provincia){
  prov<-toupper(provincia)
  b<-"http://www.axesor.es/directorio-informacion-empresas/
   empresas-de-"
```

```
p<-c("albacete","alicante","almeria","alava","asturias","
 avila","badajoz","baleares","barcelona", "vizcaya","burgos",
 "caceres","cadiz","cantabria","castellon","ciudad-real","
 cordoba","la-coruna","cuenca","guipuzcoa","girona","granada"
 ,"guadalajara","huelva","huesca","jaen","leon","lleida","
 lugo","madrid","malaga","murcia","navarra","orense","
 palencia","las-palmas","pontevedra","la-rioja","salamanca","
 santa-cruz-de-tenerife","segovia","sevilla","soria","
 tarragona","teruel","toledo","valencia","valladolid","zamora
 ","zaragoza","ceuta","melilla")
names(p)<-c("ALBACETE","ALICANTE","ALMERIA","ARABA","ASTURIAS
 ","AVILA","BADAJOZ","BALEARES","BARCELONA","BIZKAIA","BURGOS
 ","CACERES","CADIZ","CANTABRIA","CASTELLON","CIUDAD REAL","
 CORDOBA","A CORU\u00D1A","CUENCA","GIPUZKOA","GIRONA","
 GRANADA","GUADALAJARA","HUELVA","HUESCA","JAEN","LEON","
 LLEIDA","LUGO","MADRID","MALAGA","MURCIA","NAVARRA","OURENSE
 ","PALENCIA","LAS PALMAS","PONTEVEDRA", "LA RIOJA","
 SALAMANCA","TENERIFE","SEGOVIA","SEVILLA","SORIA","TARRAGONA
 ","TERUEL","TOLEDO","VALENCIA","VALLADOLID","ZAMORA","
 ZARAGOZA","CEUTA","MELILLA")
url<-paste(b,p[prov],sep="")
mun<-municipio(url)
cual<-nn.municipio(url)
display<-c(cual, paste("[",length(cual)+1,"] Todos",sep=""))
cat(display, fill=FALSE)
resp<-readline("Indique el numero del muncipio de su interes:
  ")
if ((length(cual)+1)==as.numeric(resp)){
  resp <- seq_along(cual)
} else {
  resp <- as.integer(unlist(strsplit(resp," ")))
}
set<-mun[resp]
set
lista<-unlist(sapply(set,lista.empresa,USE.NAMES=FALSE))
cat("se analizan",length(lista),"casos \n")
pp<-pbapply::pblapply(lista,empresa)
salida <- do.call(rbind.data.frame,pp)
```

```
  return(salida)
}
```

ALGORITHM A.6: empresa.a function code

```
# Extract the self-employmet's information from the Axesor
   webpage.



empresa.a<-function(http){
      pg<-read_html(http)
      nodes <- html_node(pg,css=".section-content")
      tabla <- html_table(html_children(nodes))[[1]]
      Nombre <- ifelse(length(which(tabla[,1]==paste("Aut","\
   u00ED","nomo / Profesional:",sep="")))==0,NA,tabla[which(
   tabla[,1]==paste("Aut","\u00ED","nomo / Profesional:",sep=""
   )),2])
      if(is.na(Nombre)){Nombre<-str_replace_all(html_text(html_
   node(pg,css="h3")), "([\n\r\t])", "")}
      C.A.E <- ifelse(length(which(tabla[,1]=="CNAE:"))==0,NA,
   tabla[which(tabla[,1]=="CNAE:"),2])
      if(!is.na(C.A.E)){
        CodCAE<-na.omit(unlist(strsplit(C.A.E,"[^[:digit:]]")))
   [1]
        DescCAE<-sub("[[:digit:]]+","",str_replace_all(C.A.E, "
   ([\n\r\t])", ""))
        if(length(DescCAE)==0){ DescCAE<-NA}
      } else {
        CodCAE<-NA
        DescCAE<-NA
      }
      S.I.C <- ifelse(length(which(tabla[,1]=="SIC:"))==0,NA,
   tabla[which(tabla[,1]=="SIC:"),2])
      if(!is.na(S.I.C)){
        CodSIC<-na.omit(unlist(strsplit(S.I.C,"[^[:digit:]]")))
   [1]
```

```
      DescSIC<-sub("[[:digit:]]+","",str_replace_all(S.I.C, "
([\n\r\t])", ""))
      if(length(DescSIC)==0){ DescSIC<-NA}
    } else {
      CodSIC<-NA
      DescSIC<-NA
    }
    Dir <- ifelse(length(which(tabla[,1]==paste("Direcci","\
u00F3","n:",sep=""))==0,NA,tabla[which(tabla[,1]==paste("
Direcci","\u00F3","n:",sep="")),2])
    if(!is.na(Dir)){
      Dd <- html_children(html_node(nodes,"span"))
      if(length(Dd)==0){Dd <- html_nodes(nodes,"span")}
      Direccion <- str_trim(str_replace_all(html_text(Dd[1]),
 "([\n\r\t,])", ""))
      Cod_postal <- str_trim(str_replace_all(html_text(Dd)
[2], "([\n\r\t,])", ""))
      Mun <- simpleCap(str_trim(str_replace_all(html_text(Dd)
[3], "([\n\r\t,])", "")))
      Prov <- simpleCap(str_trim(str_replace_all(html_text(Dd
)[4], "([\n\r\t,])", "")))
    } else {
      Direccion <- NA
      Cod_postal <- NA
      Mun <- NA
      Prov <- NA
    }
    p2<-html_node(pg,css="#resumen_general")
    data<-html_text(html_nodes(p2,css="p")[2])
    geo<-html_nodes(html_nodes(p2,css="div"),css="span")
    lat<-as.numeric(html_text(geo[grep("latitude",capture.
output(geo))-1]))
    lng<-as.numeric(html_text(geo[grep("longitude",capture.
output(geo))-1]))
    if (sum(lat)==0){lat<-0 ; lng<-0}
    web_aexor<-http
          fila<-data.frame("Provinciaa"=Prov,
            "Municipalidad"=Mun,
```

```
                "Nombre"=Nombre,
                "Direccion"=Direccion,
                "Codigo Postal"=Cod_postal,
                "C.N.A.E"=CodCAE,
                "Descripcion C.N.A.E"=DescCAE,
                "S.I.C"=CodSIC,
                "Descripcion S.I.C"=DescSIC,
                "Latitud"=lat,
                "Longitud"=lng,
                "URL en Axesor"=web_aexor,stringsAsFactors =
   FALSE)
      return(fila)
}
```

ALGORITHM A.7: empresa function code

```
# Extract the firm's information from the Axesor webpage.



empresa<-function(http){
      pg<-read_html(http)
      nodes <- html_node(pg,css=".section-content")
      tabla <- html_table(html_children(nodes))[[1]]
      Nombre <- ifelse(length(which(tabla[,1]=="Nombre:"))==0,
   NA,tabla[which(tabla[,1]=="Nombre:"),2])
      if(is.na(Nombre)){Nombre<-str_replace_all(html_text(html_
   node(pg,css="h3")), "([\n\r\t])", "")}
      CIF <- ifelse(length(which(tabla[,1]=="CIF:"))==0,NA,
   tabla[which(tabla[,1]=="CIF:"),2])
      Forma_juridica <- ifelse(length(which(tabla[,1]==paste("
   Forma jur","\u00ED","dica:",sep=""))==0,NA,tabla[which(
   tabla[,1]==paste("Forma jur","\u00ED","dica:",sep=""))),2])
      Nace <- ifelse(length(which(tabla[,1]=="Constituida hace:
   "))==0,"N.I",tabla[which(tabla[,1]=="Constituida hace:"),2])
      Objetivo_social <- ifelse(length(which(tabla[,1]=="Objeto
    social:"))==0,"N.I",str_replace_all(tabla[which(tabla[,1]==
   "Objeto social:"),2], "([\n\r\t])", ""))
```

```
   C.A.E <- ifelse(length(which(tabla[,1]=="CNAE:"))==0,NA,
tabla[which(tabla[,1]=="CNAE:"),2])
   if(!is.na(C.A.E)){
     CodCAE<-na.omit(unlist(strsplit(C.A.E,"[^[:digit:]]")))
[1]
     DescCAE<-sub("[[:digit:]]+","",str_replace_all(C.A.E, "
([\n\r\t])", ""))
     if(length(DescCAE)==0){ DescCAE<-NA}
   } else {
     CodCAE<-NA
     DescCAE<-NA
   }
   S.I.C <- ifelse(length(which(tabla[,1]=="SIC:"))==0,NA,
tabla[which(tabla[,1]=="SIC:"),2])
   if(!is.na(S.I.C)){
     CodSIC<-na.omit(unlist(strsplit(S.I.C,"[^[:digit:]]")))
[1]
     DescSIC<-sub("[[:digit:]]+","",str_replace_all(S.I.C, "
([\n\r\t])", ""))
     if(length(DescSIC)==0){ DescSIC<-NA}
   } else {
     CodSIC<-NA
     DescSIC<-NA
   }
   Dir <- ifelse(length(which(tabla[,1]==paste("Direcci","\
u00F3","n:",sep="")))==0,NA,tabla[which(tabla[,1]==paste("
Direcci","\u00F3","n:",sep="")),2])
   if(!is.na(Dir)){
     Dd <- html_children(html_node(nodes,"span"))
     if(length(Dd)==0){Dd <- html_nodes(nodes,"span")}
     Direccion <- str_trim(str_replace_all(html_text(Dd[1]),
 "([\n\r\t,])", ""))
     Cod_postal <- str_trim(str_replace_all(html_text(Dd)
[2], "([\n\r\t,])", ""))
     Mun <- simpleCap(str_trim(str_replace_all(html_text(Dd)
[3], "([\n\r\t,])", "")))
     Prov <- simpleCap(str_trim(str_replace_all(html_text(Dd
)[4], "([\n\r\t,])", "")))
```

```
    } else {
      Direccion <- NA
      Cod_postal <- NA
      Mun <- NA
      Prov <- NA
    }
    p2<-html_node(pg,css="#resumen_general")
    data <- html_text(html_children(p2)[3])
    Tramo_cap_social<-str_match(data,paste("en el tramo de ",
"(.+)","\u20AC,",sep=""))[,2]
    Tramo_empleados<-sub(" y ","-",str_match(data,paste("
empleados de entre ","(.+)"," y un importe",sep=""))[,2])
    Tramo_ventas<-sub(" y ","-",str_match(data,paste("ventas
de entre ","(.+)","\u20AC.",sep=""))[,2])
    ss<-unlist(strsplit(Tramo_ventas,"-"))
    M_c_ventas<-(as.numeric(gsub("[.]","",ss[1]))+as.numeric(
gsub("[.]","",ss[2])))/2
    ee<-unlist(strsplit(Tramo_empleados,"-"))
    M_c_empleados<-(as.numeric(gsub("[.]","",ee[1]))+as.
numeric(gsub("[.]","",ee[2])))/2
    cc<-sapply(unlist(strsplit(Tramo_cap_social,"-")),str_
trim,USE.NAMES = FALSE)
    M_c_cap_social<-(as.numeric(gsub("[.]","",cc[1]))+as.
numeric(gsub("[.]","",cc[2])))/2
    geo<-html_nodes(html_nodes(p2,css="div"),css="span")
    lat<-as.numeric(html_text(geo[grep("latitude",capture.
output(geo))-1]))
    lng<-as.numeric(html_text(geo[grep("longitude",capture.
output(geo))-1]))
    if (sum(lat)==0){lat<-0 ; lng<-0}
    web_aexor<-http
    fila<-data.frame("Provincia"=Prov,
            "Municipalidad"=Mun,
            "Nombre"=Nombre,
            "CIF"=CIF,
            "Forma Juridica"=Forma_juridica,
            "Constituida Hace"=Nace,
            "Objeto Social"= Objetivo_social,
```

```
                "Direccion"=Direccion,
                "Codigo Postal"=Cod_postal,
                "C.N.A.E"=CodCAE,
                "Descripcion C.N.A.E"=DescCAE,
                "S.I.C"=CodSIC,
                "Descripcion S.I.C"=DescSIC,
                "Tramo Capital Social"=Tramo_cap_social,
                "Tramo Empleados"=Tramo_empleados,
                "Tramo Ventas"=Tramo_ventas,
                "Marca de clase Cap. Social"=M_c_cap_social,
                "Marca de clases empleados"=M_c_empleados,
                "Marca de clase ventas"=M_c_ventas,
                "Latitud"=lat,
                "Longitud"=lng,
                "URL en Axesor"=web_aexor,stringsAsFactors =
   FALSE)
      return(fila)
}
```

Algorithm A.8: get,empresa function code

```
# Crea las URLs de las empresas para extraer informacion


get.empresas<-function(http){
      h<-read_html(http)
      l.e<-html_node(h,css="table")
      g<-html_children(html_nodes(html_node(l.e,css="tbody"),
   css="td"))
      e<-html_attr(g,"href")
      e<-e[complete.cases(e)==TRUE]
      www<-rep("http:",length(e))
      w.e<-paste(www,e,sep="")
      w.e
    }
```

Algorithm A.9: get.empresas.a function code

```
# Crea las URLs de los Autonomos para extraer informacion


get.empresas.a<-function(http){
      h<-read_html(http)
      l.e<-html_node(h,css="table")
      g<-html_children(html_nodes(html_node(l.e,css="tbody"),
   css="td"))
      e<-html_attr(g,"href")
      e<-e[complete.cases(e)==TRUE]
      www<-rep("https://autonomos.axesor.es/",length(e))
      w.e<-paste(www,e,sep="")
      w.e
    }
```

ALGORITHM A.10: getbase.fen function code

```
#' @importFrom utils download.file capture.output read.table
#' @import stringr
#' @import xlsx
#' @import rvest
#' @importFrom xml2 read_html
#' @importFrom stats na.omit
#' @importFrom XML htmlParse
#' @importFrom stringi stri_count
#'
#' @name getbase.fen
#' @rdname getbase.fen
#'
#' @title Collect information of municipality population
   phenomena.
#'
#' @description \code{getbase.fen} downloads the population
   phenomena data of the provinces municipalities
   corresponding to the required.
#'
#' @param year A numerical value between 1996 and the current
   year indicating the year of the required data.
#' @param provincia one of the 52 Spains province.
```

```
#'
#' @family download functions
#'
#' @details The names of the Spanish provinces may or may not
   be called. If yes, employ capital letters as follows:
#'
#' "ALBACETE", "ALICANTE", "ALMERIA", "ARABA", "ASTURIAS", "
   AVILA", "BADAJOZ", "BALEARES", "BARCELONA", "BIZKAIA", "
   BURGOS", "CACERES", "CADIZ", "CANTABRIA", "CASTELLON", "
   CIUDAD REAL", "CORDOBA", "A CORUÑA", "CUENCA", "GIPUZKOA", "
   GIRONA", "GRANADA", "GUADALAJARA", "HUELVA", "HUESCA", "JAEN
   ", "LEON", "LLEIDA", "LUGO", "MADRID", "MALAGA", "MURCIA", "
   NAVARRA", "OURENSE", "PALENCIA",  "LAS PALMAS", "PONTEVEDRA
   ", "LA RIOJA", "SALAMANCA", "TENERIFE", "SEGOVIA", "SEVILLA
   ","SORIA", "TARRAGONA", "TERUEL", "TOLEDO", "VALENCIA", "
   VALLADOLID", "ZAMORA", "ZARAGOZA", "CEUTA" and "MELILLA"
#'
#' @return It is a \code{xlsx} which is host in \code{data_
   poblacion} folder iwhich, in turn, is located inside the
   working directory called \code{fen_year_provincia}.
#' @examples
#' getbase.fen(2005,"Madrid")
#'
#' @export


getbase.fen<-function(year,provincia){
    year<-as.character(year)
    provincia<-toupper(provincia)
    provincia<-a.letter(provincia)
    if(dir.exists(file.path(getwd(),"data_poblacion"))==FALSE){
    dir.create(file.path(getwd(),"data_poblacion"))
    }
    file<-paste(getwd(),"/data_poblacion/",paste("fen",year,
   provincia,sep="_"),".xlsx",sep="")
```

```
    p<-c("02","03","04","01","33","05","06","07","08","48","09"
   ,"10","11","39","12","13","14","15","16","20","17","18","19"
   ,"21","22","23","24","25","27","28","29","30","31","32","34"
   ,"35","36","26","37","38","40","41","42","43","44","45","46"
   ,"47","49","50","51","52")
    names(p) <- c("ALBACETE","ALICANTE","ALMERIA","ARABA","
   ASTURIAS","AVILA","BADAJOZ","BALEARES","BARCELONA","BIZKAIA"
   ,"BURGOS","CACERES","CADIZ","CANTABRIA","CASTELLON","CIUDAD_
   REAL","CORDOBA","A_CORUNA","CUENCA","GIPUZKOA","GIRONA","
   GRANADA","GUADALAJARA","HUELVA","HUESCA","JAEN","LEON","
   LLEIDA","LUGO","MADRID","MALAGA","MURCIA","NAVARRA","OURENSE
   ","PALENCIA","LAS_PALMAS","PONTEVEDRA","LA_RIOJA","SALAMANCA
   ","TENERIFE","SEGOVIA","SEVILLA","SORIA","TARRAGONA","TERUEL
   ","TOLEDO","VALENCIA","VALLADOLID","ZAMORA","ZARAGOZA","
   CEUTA","MELILLA")
    n<-p[provincia]
    url<-paste("http://www.ine.es/jaxi/files/_px/es/xlsx/t20/
   e301/fenom/a",year,"/l0/","230",n,".px?nocab=1",sep="")
    download.file(url,file, mode='wb')
}
```

Algorithm A.11: getbase.paro function code

```
#' @name getbase.paro
#' @rdname getbase.paro
#'
#' @title Collects information of unemployed at a municipality
   level.
#'
#' @description \code{getbase.paro} downloads data about the
   unemployed of the Spanish municipalities by province.
#'
#' @param year a numerical value between 2005 and the last
   available, which indicates the year of the required database
   .
#' @param mes oone of the 12 months  in  Spanish language
   indicating the month of the data collection.
#' @param provincia one of the 52 Spainish provinces.
#'
```

```
#' @return It is a \code{xlsx} file host in the \code{data_paro
   } folder which, in turn, is located inside the working
   directory called  \code{paro_MUNI_provincia_mmyy.xls}.
#'
#' @examples
#' getbase.paro(2005,"julio","Madrid")
#'
#' @details The month must be called as follows: "enero", "
   febrero", "marzo", "abril", "mayo", "junio", "julio", "
   agosto", "septiembre", "octubre", "noviembre" and "diciembre
   ".
#'
#'  The names of the Spanish provinces may or may not be called
   . If yes, employ capital letters as follows:
#'
#' "ALBACETE", "ALICANTE", "ALMERIA", "ARABA", "ASTURIAS", "
   AVILA", "BADAJOZ", " BALEARES", "BARCELONA", "BIZKAIA", "
   BURGOS", "CACERES", "CADIZ", "CANTABRIA", "CASTELLON", "
   CIUDAD REAL", "CORDOBA", "A CORUÑA", "CUENCA", "GIPUZKOA", "
   GIRONA", "GRANADA", "GUADALAJARA", "HUELVA", "HUESCA", "JAEN
   ", "LEON", "LLEIDA", "LUGO", "MADRID", "MALAGA", "MURCIA", "
   NAVARRA", "OURENSE", "PALENCIA",  "LAS PALMAS", "PONTEVEDRA
   ", "LA RIOJA", "SALAMANCA", "TENERIFE", "SEGOVIA", "SEVILLA
   ","SORIA", "TARRAGONA", "TERUEL", "TOLEDO", "VALENCIA", "
   VALLADOLID", "ZAMORA", "ZARAGOZA", "CEUTA" and "MELILLA"
#'
#' @family download functions
#' @export

getbase.paro<-function(year,mes,provincia){
  year<-as.character(year)
  if(dir.exists(file.path(getwd(),"data_paro"))==FALSE){
    dir.create(file.path(getwd(),"data_paro"))
    }
  provincia<-toupper(provincia)
  provincia<-a.letter(provincia)
  if(as.numeric(year)<=2011){
    if(provincia=="ARABA"){provincia <- "ALAVA"}
```

```
    if(provincia=="BIZKAIA"){provincia <- "VIZCAYA"}
    if(provincia=="GIPUZKOA"){provincia <- "GUIPUZCOA"}
    }
  mes<-tolower(mes)
  nn.mes<-seq(1,12,1)
  names(nn.mes)<-c("enero","febrero","marzo","abril","mayo","
   junio","julio","agosto","septiembre","octubre","noviembre","
   diciembre")
  cod<-paste("0",nn.mes[mes],substr(year,3,4),sep="")
  name<-paste(paste("MUNI",provincia,cod,sep="_"),".xls",sep=""
   )
  url<-paste("http://www.sepe.es/contenidos/que_es_el_sepe/
   estadisticas/datos_estadisticos/municipios/",year,"/",paste(
   mes,year,sep="_"),"/",name,sep="")
  dir<-paste(getwd(),"/data_paro/",sep="")
  file<-paste(dir,"paro_",name,sep="")
  download.file(url,file, mode='wb')
}
```

Algorithm A.12: getbase.pob function code

```
#' @name getbase.pob
#' @rdname getbase.pob
#'
#' @title Collects information of population at a municipality
   level.
#'
#' @description \code{getbase.pob} downloads population data of
    the Spanish municipalities by province.
#'
#' @param year a numerical value from 1996 and the latest
   available, which indicates the year of the required database
   .
#' @param provincia one of the 52 Spainish province.
#' @param extr logical variable  is  it foreign population?
    FALSE is the default value.
#' @param anual logical variable  is  data required by age?
    FALSE is the default value.
#'
```

```
#' @return It is a \code{xlsx} which is host in  the \code{data
   _poblacion} which, in turn, is located inside the working
   directory called \code{pob_q_year_provincia.xls}. If \code{
   extr} is TRUE, the file is called \code{pob_e_year_provincia
   .xls}. In case of \code{anual} is TRUE the file is called \
   code{pob_a_year_provincia.xls}. There is not data for
   foreign population by one or five-year age, the combination
   of \code{extr=TRUE} and \code{anual=TRUE will} generate an
   error message (  No   data for these  c a s e s ).
#'
#' @details The names of the Spanish provinces may or may not
   be called. If yes, employ capital letters as follows:
#'
#' "ALBACETE", "ALICANTE", "ALMERIA", "ARABA", "ASTURIAS", "
   AVILA", "BADAJOZ", "BALEARES", "BARCELONA", "BIZKAIA", "
   BURGOS", "CACERES", "CADIZ", "CANTABRIA", "CASTELLON", "
   CIUDAD REAL", "CORDOBA", "A CORUÑA", "CUENCA", "GIPUZKOA", "
   GIRONA", "GRANADA", "GUADALAJARA", "HUELVA", "HUESCA", "JAEN
   ", "LEON", "LLEIDA", "LUGO", "MADRID", "MALAGA", "MURCIA", "
   NAVARRA", "OURENSE", "PALENCIA",  "LAS PALMAS", "PONTEVEDRA
   ", "LA RIOJA", "SALAMANCA", "TENERIFE", "SEGOVIA", "SEVILLA
   ","SORIA", "TARRAGONA", "TERUEL", "TOLEDO", "VALENCIA", "
   VALLADOLID", "ZAMORA", "ZARAGOZA", "CEUTA" and "MELILLA"
#'
#' @family download functions
#' @examples
#' getbase.pob(2005,"Madrid")
#'
#' @export


getbase.pob<-function(year,provincia,extr=FALSE,anual=FALSE){
    year<-as.character(year)
    if(extr==TRUE && anual==TRUE) stop("No data for these cases
   ")
    if(anual==TRUE && as.numeric(year)<2011) stop("No data for
   these cases")
    provincia<-toupper(provincia)
```

```
    provincia<-a.letter(provincia)
    if(dir.exists(file.path(getwd(),"data_poblacion"))==FALSE){
    dir.create(file.path(getwd(),"data_poblacion"))
    }
    if (extr==TRUE) {
      if(as.numeric(year)<2002) {q<-4} else {q<-2}
      e<-"e"
      } else {
      q<-1
      e<-"q"
      }
    if (anual==TRUE) { q<-6; e<-"a"}
    file<-paste(getwd(),"/data_poblacion/",paste("pob",e,year,
   provincia,sep="_"),".xlsx",sep="")
    p<-c("02","03","04","01","33","05","06","07","08","48","09"
   ,"10","11","39","12","13","14","15","16","20","17","18","19"
   ,"21","22","23","24","25","27","28","29","30","31","32","34"
   ,"35","36","26","37","38","40","41","42","43","44","45","46"
   ,"47","49","50","51","52")
    names(p) <- c("ALBACETE","ALICANTE","ALMERIA","ARABA","
   ASTURIAS","AVILA","BADAJOZ","BALEARES","BARCELONA","BIZKAIA"
   ,"BURGOS","CACERES","CADIZ","CANTABRIA","CASTELLON","CIUDAD_
   REAL","CORDOBA","A_CORUNA","CUENCA","GIPUZKOA","GIRONA","
   GRANADA","GUADALAJARA","HUELVA","HUESCA","JAEN","LEON","
   LLEIDA","LUGO","MADRID","MALAGA","MURCIA","NAVARRA","OURENSE
   ","PALENCIA","LAS_PALMAS","PONTEVEDRA","LA_RIOJA","SALAMANCA
   ","TENERIFE","SEGOVIA","SEVILLA","SORIA","TARRAGONA","TERUEL
   ","TOLEDO","VALENCIA","VALLADOLID","ZAMORA","ZARAGOZA","
   CEUTA","MELILLA")
    n<-p[provincia]
    url<-paste("http://www.ine.es/jaxi/files/_px/es/xlsx/t20/
   e245/p05/a",year,"/l0/000",n,"00",q,".px?nocab=1",sep="")
    print(url)
    download.file(url,file,, mode='wb')
}
```

Algorithm A.13: growth function code

```
# Compute the groth rate of n periods od an x variable

growth<-function(x,y,n){
  z<-(((x/y)^(1/n))-1)*100
  z
  }
```

ALGORITHM A.14: ind.ev function code

```
#' @name ind.ev
#' @rdname ind.ev
#'
#' @title Panel of demographic indexes at a municipality level
   for a period of time
#'
#' @description \code{ind.ev} creates a list containing a panel
    of demographic indexes of Spain at the municipality level
   for a period of time from the years \code{inicio} to \code{
   fin}.
#' @param inicio starting year of the panel, which must be
   higher than 1996.
#' @param fin last year of the panel.
#' @param provincia one of the 52 Spanish provinces.
#' @param print logical variable   do   you need to print a
   output file with the results?   , being FALSE the default
   value
#'
#' @return It is a list conteining ten data frame.
#'
#' @details This function calculates ten demographical indexes:
#' \itemize{
#'  \item Childhood index
#'  \item Youth index
#'  \item Third age index
#'  \item Dependence index
#'  \item Unemployment rate, both sexes
#'  \item Unemployment rate, males
#'  \item Unemployment rate, females
#'  \item Municipality average age, both sexes
```

```
#'  \item Municipality average age, males
#'  \item Municipality average age, females
#' }
#' For a full description of the index, see the 2017
   Socioeconomic Atlas of Extremadura
#' If \code{print} is set to \code{TRUE}, an \code{xlsx} file
   containing ten sheet -one  data frame per index- is saved
   into the folder \code{Outputs} which the name  \code{pob_ev_
   index_provincia_inicio-fin.xlsx}
#'
#' @references{
#' Junta de Extremadura (2017) Atlas Socieconómico de
   Extremadura 2017. Mérida (Spain).
#' \url{http://estadistica.gobex.es/web/guest/atlas-
   socieconomico-de-extremadura}
#'}
#' @family manipulation functions
#' @examples
#' ind.ev(2005,2007,"Avila")
#'
#' @export


ind.ev<-function(inicio,fin,provincia,print=FALSE){
  if(fin<inicio) stop("La fecha de inicio debe ser mayor que la
    fecha de fin \n")
  n<-seq(inicio,fin,1)
  year<-as.character(sort(n,decreasing=TRUE))
  entrada<-pob.ind(year[1],provincia)
  t<-dim(entrada)[1]
  base.i<-cbind(entrada[,c(1,2,3)])
  base.j<-cbind(entrada[,c(1,2,4)])
  base.v<-cbind(entrada[,c(1,2,5)])
  base.d<-cbind(entrada[,c(1,2,6)])
  base.p<-cbind(entrada[,c(1,2,7)])
  base.ph<-cbind(entrada[,c(1,2,8)])
  base.pm<-cbind(entrada[,c(1,2,9)])
  base.m<-cbind(entrada[,c(1,2,10)])
```

```
base.mh<-cbind(entrada[,c(1,2,11)])
base.mm<-cbind(entrada[,c(1,2,12)])
  for (i in 2:length(year)){
  aux.i<-rep(NA,t)
  aux.j<-rep(NA,t)
  aux.v<-rep(NA,t)
  aux.d<-rep(NA,t)
  aux.p<-rep(NA,t)
  aux.ph<-rep(NA,t)
  aux.pm<-rep(NA,t)
  aux.m<-rep(NA,t)
  aux.mh<-rep(NA,t)
  aux.mm<-rep(NA,t)
  act<-pob.ind(year[i],provincia)
  v<-intersect(entrada[,1],act[,1])
  for(j in 1:length(v)){
    b<-which(base.i[,1]==v[j])
    e<-which(act[,1]==v[j])
    aux.i[b]<-act[e,3]
    aux.j[b]<-act[e,4]
    aux.v[b]<-act[e,5]
    aux.d[b]<-act[e,6]
    aux.p[b]<-act[e,7]
    aux.ph[b]<-act[e,8]
    aux.pm[b]<-act[e,9]
    aux.m[b]<-act[e,10]
    aux.mh[b]<-act[e,11]
    aux.mm[b]<-act[e,12]
  }
  base.i<-cbind(base.i,aux.i)
  base.j<-cbind(base.j,aux.j)
  base.v<-cbind(base.v,aux.v)
  base.d<-cbind(base.d,aux.d)
  base.p<-cbind(base.p,aux.p)
  base.ph<-cbind(base.ph,aux.ph)
  base.pm<-cbind(base.pm,aux.pm)
  base.m<-cbind(base.m,aux.m)
  base.mh<-cbind(base.mh,aux.mh)
```

```
    base.mm<-cbind(base.mm,aux.mm)
    }
  colnames(base.i)<-c("Cod","Municipio",year)
  colnames(base.j)<-c("Cod","Municipio",year)
  colnames(base.v)<-c("Cod","Municipio",year)
  colnames(base.d)<-c("Cod","Municipio",year)
  colnames(base.p)<-c("Cod","Municipio",year)
  colnames(base.ph)<-c("Cod","Municipio",year)
  colnames(base.pm)<-c("Cod","Municipio",year)
  colnames(base.m)<-c("Cod","Municipio",year)
  colnames(base.mh)<-c("Cod","Municipio",year)
  colnames(base.mm)<-c("Cod","Municipio",year)
  orden<-c(1,2,seq(dim(base.i)[2],3))
  base.i<-base.i[,orden]
  base.j<-base.j[,orden]
  base.v<-base.v[,orden]
  base.d<-base.d[,orden]
  base.p<-base.p[,orden]
  base.ph<-base.ph[,orden]
  base.pm<-base.pm[,orden]
  base.m<-base.m[,orden]
  base.mh<-base.mh[,orden]
  base.mm<-base.mm[,orden]
  out<-list(base.i,base.j,base.v,base.d,base.p,base.ph,base.pm,
   base.m,base.mh,base.mm)
  names(out)<-c("Infancia","Juventud","Vejez","Dependencia","
   Paro Total","Paro Hombres","Paro Mujeres","Edad Media","Edad
    Media Hombres","Edad Media Mujeres")
  if(print==TRUE){
    excel<-createWorkbook()
    s1<-createSheet(excel,sheetName="Infancia")
    s2<-createSheet(excel,sheetName="Juventud")
    s3<-createSheet(excel,sheetName="Vejez")
    s4<-createSheet(excel,sheetName="Dependencia")
    s5<-createSheet(excel,sheetName="Paro Total")
    s6<-createSheet(excel,sheetName="Paro Hombres")
    s7<-createSheet(excel,sheetName="Paro Mujeres")
    s8<-createSheet(excel,sheetName="Edad Media")
```

```
    s9<-createSheet(excel,sheetName="Edad Media Hombres")
    s10<-createSheet(excel,sheetName="Edad Media Mujeres")
    addDataFrame(base.i,s1)
    addDataFrame(base.j,s2)
    addDataFrame(base.v,s3)
    addDataFrame(base.d,s4)
    addDataFrame(base.p,s5)
    addDataFrame(base.ph,s6)
    addDataFrame(base.pm,s7)
    addDataFrame(base.m,s8)
    addDataFrame(base.mh,s9)
    addDataFrame(base.mm,s10)
      if(dir.exists(file.path(getwd(),"Outputs"))==FALSE){
      dir.create(file.path(getwd(),"Outputs"))
      }
    file<-paste(getwd(),"/Output/pob_ev_index_",provincia,"_",
   paste(inicio,fin,sep="-"),".xlsx",sep="")
    saveWorkbook(excel,file)
  }
  out
  }
```

Algorithm A.15: lista.empresa.a function code

```
# Generate a list containing the self-employment's URLs in
   Axesor web site


lista.empresa.a<-function(http){
  h<-read_html(http)
  doc<-htmlParse(h)
  txt<-capture.output(doc)
  v<-grep("paginacion-botones",txt)
    if (sum(v)==0){
      salida<-get.empresas(http)
    } else {
      n.p<-html_attr(html_children(html_node(h,css=".paginacion
   -numeracion")),"href")
```

```
      pag<-as.numeric(str_extract(n.p[length(n.p)],"[[:digit
   :]]+"))
      times<-as.character(seq(1,pag))
      p.g1<-rep(substr(http,1,nchar(http)-1),length(times))
      paginas<-paste(p.g1,times,sep="")
      salida<-lapply(paginas,get.empresas.a)
      salida<-unlist(salida,use.names=FALSE)
      }
    salida
}
```

ALGORITHM A.16: lista.empresa function code

```
# Generate a list containing the Firms URLs in Axesor web
   site



lista.empresa<-function(http){
  h<-read_html(http)
  doc<-htmlParse(h)
  txt<-capture.output(doc)
  v<-grep("paginacion-botones",txt)
    if (sum(v)==0){
      salida<-get.empresas(http)
    } else {
      n.p<-html_attr(html_children(html_node(h,css=".paginacion
   -numeracion")),"href")
      pag<-as.numeric(str_extract(n.p[length(n.p)],"[[:digit
   :]]+"))
      times<-as.character(seq(1,pag))
      p.g1<-rep(substr(http,1,nchar(http)-1),length(times))
      paginas<-paste(p.g1,times,sep="")
      salida<-lapply(paginas,get.empresas)
      salida<-unlist(salida,use.names=FALSE)
      }
    salida
}
```

ALGORITHM A.17: municipio.a function code

```
# Generate a list muncipalities URLs in Axesor self-employment
   data base


municipio.a<-function(http){
    base.m<-html_nodes(read_html(http),css="#bloque_
   listadoMunicipios")
    l.m<-html_attr(html_children(html_nodes(base.m,css="td")),"
   href")
    mun<-rep(NA,length(l.m))
    for (j in 1:length(l.m)){
      mun[j]<-paste("https://autonomos.axesor.es/",l.m[j],sep="
   ")
    }
    return(mun)
  }
```

ALGORITHM A.18: municipio function code

```
# Generate a list muncipalities URLs in Axesor firms data base


municipio<-function(http){
    base.m<-html_nodes(read_html(http),css="#bloque_
   listadoMunicipios")
    l.m<-html_attr(html_children(html_nodes(base.m,css="td")),"
   href")
    mun<-rep(NA,length(l.m))
    for (j in 1:length(l.m)){
      mun[j]<-paste("http:",l.m[j],sep="")
    }
    return(mun)
  }
```

ALGORITHM A.19: nn.municipio function code

```
# Genera una lista numerada de los municipios

nn.municipio<-function(http){
    base.m<-html_nodes(read_html(http),css="#bloque_
   listadoMunicipios")
    n.m<-html_text(html_children(html_nodes(base.m,css="td")),"
   href")
    id<-seq(1,length(n.m),1)
    names<-paste("[",id,"] ",n.m,"\n",sep="")
    names
  }
```

ALGORITHM A.20: num.firm.a function code

```
#' @name num.firm.a
#' @rdname num.firm.a
#'
#' @title Collects the set of self-employed at a municipality
   level
#'
#' @description \code{num.firm.a} generates a data frame with
   the set of self-employed in a selected municipality
#'
#' @param provincia one of the 52 Spanish provinces. See \link{
   getbase.pob} for details.
#'
#'
#'
#' @return It is a data frame
#'
#' @family firm functions
#' @examples
#' \dontrun{num.firm.a("Araba")}
#'
#' @export
```

```
num.firm.a <- function(provincia){
  prov <- toupper(provincia)
  b <- "https://autonomos.axesor.es/informe-de-autonomo/
   provincias/"
  p <- c("albacete","alicante","almeria","alava","asturias","
   avila","badajoz","baleares","barcelona", "vizcaya","burgos",
   "caceres","cadiz","cantabria","castellon","ciudad-real","
   cordoba","la-coruna","cuenca","guipuzcoa","girona","granada"
   ,"guadalajara","huelva","huesca","jaen","leon","lleida","
   lugo","madrid","malaga","murcia","navarra","orense","
   palencia","las-palmas","pontevedra","la-rioja","salamanca","
   santa-cruz-de-tenerife","segovia","sevilla","soria","
   tarragona","teruel","toledo","valencia","valladolid","zamora
   ","zaragoza","ceuta","melilla")
  names(p) <- c("ALBACETE","ALICANTE","ALMERIA","ARABA","
   ASTURIAS","AVILA","BADAJOZ","BALEARES","BARCELONA","BIZKAIA"
   ,"BURGOS","CACERES","CADIZ","CANTABRIA","CASTELLON","CIUDAD
   REAL","CORDOBA","A CORU\u00D1A","CUENCA","GIPUZKOA","GIRONA"
   ,"GRANADA","GUADALAJARA","HUELVA","HUESCA","JAEN","LEON","
   LLEIDA","LUGO","MADRID","MALAGA","MURCIA","NAVARRA","OURENSE
   ","PALENCIA","LAS PALMAS","PONTEVEDRA", "LA RIOJA","
   SALAMANCA","TENERIFE","SEGOVIA","SEVILLA","SORIA","TARRAGONA
   ","TERUEL","TOLEDO","VALENCIA","VALLADOLID","ZAMORA","
   ZARAGOZA","CEUTA","MELILLA")
  url <- paste(b,p[prov],sep="")
  mun <- municipio.a(url)
  aux <- sapply(mun,lista.empresa.a)
  n <- sapply(aux,length)
  names(n) <- NULL
  names.aux <- html_nodes(read_html(url),css="#bloque_
   listadoMunicipios")
  names <- html_text(html_children(html_nodes(names.aux,css="td
   ")),"href")
  output <- data.frame("Municipality"=names, "N firms"=n)
  return(output)
}
```

Algorithm A.21: num.firm function code

```
#' @name num.firm
#' @rdname num.firm
#'
#' @title Collects the set of firms at a municipality level
#'
#' @description \code{num.firm} generates a data frame with the
    set of firms in a selected municipality
#'
#' @param provincia one of the 52 Spanish provinces. See \link{
   getbase.pob} for details.
#'
#'
#' @return It is a data frame
#'
#' @family firm functions
#' @examples
#' \dontrun{num.firm("Araba")}
#'
#' @export



num.firm <- function(provincia){
  prov <- toupper(provincia)
  b <- "http://www.axesor.es/directorio-informacion-empresas/
   empresas-de-"
  p <- c("albacete","alicante","almeria","alava","asturias","
   avila","badajoz","baleares","barcelona", "vizcaya","burgos",
   "caceres","cadiz","cantabria","castellon","ciudad-real","
   cordoba","la-coruna","cuenca","guipuzcoa","girona","granada"
   ,"guadalajara","huelva","huesca","jaen","leon","lleida","
   lugo","madrid","malaga","murcia","navarra","orense","
   palencia","las-palmas","pontevedra","la-rioja","salamanca","
   santa-cruz-de-tenerife","segovia","sevilla","soria","
   tarragona","teruel","toledo","valencia","valladolid","zamora
   ","zaragoza","ceuta","melilla")
```

```
  names(p) <- c("ALBACETE","ALICANTE","ALMERIA","ARABA","
   ASTURIAS","AVILA","BADAJOZ","BALEARES","BARCELONA","BIZKAIA"
   ,"BURGOS","CACERES","CADIZ","CANTABRIA","CASTELLO","CIUDAD
   REAL","CORDOBA","A CORU\u00D1A","CUENCA","GIPUZKOA","GIRONA"
   ,"GRANADA","GUADALAJARA","HUELVA","HUESCA","JAEN","LEON","
   LLEIDA","LUGO","MADRID","MALAGA","MURCIA","NAVARRA","OURENSE
   ","PALENCIA","LAS PALMAS","PONTEVEDRA", "LA RIOJA","
   SALAMANCA","TENERIFE","SEGOVIA","SEVILLA","SORIA","TARRAGONA
   ","TERUEL","TOLEDO","VALENCIA","VALLADOLID","ZAMORA","
   ZARAGOZA","CEUTA","MELILLA")
  url <- paste(b,p[prov],sep="")
  mun <- municipio(url)
  aux <- sapply(mun,lista.empresa)
  n <- sapply(aux,length)
  names(n) <- NULL
  names.aux <- html_nodes(read_html(url),css="#bloque_
   listadoMunicipios")
  names <- html_text(html_children(html_nodes(names.aux,css="td
   ")),"href")
  output <- data.frame("Municipality"=names, "N firms"=n)
  return(output)
}
```

Algorithm A.22: paro function code

```
#' @name paro
#' @rdname paro
#'
#' @title Imports into R data of unemployed at a municipality
   level..
#'
#' @description \code{paro} imports the unemployed of the
   Spanish municipalities by province.
#'
#' @param year a numerical value from 2005 and the latest
   available, which indicates the year of the required database
   .
```

```
#' @param mes one of the 12 months  in  Spanish language
   indicating the month of the data collection. See \link{
   getbase.paro} for details.
#' @param provincia one of the 52 Spains province.
#'
#' @return It is a data frame containing the fallowing
   variables:.
#' - \code{cod} is the municipality identification number based
    in the INE codification
#' - \code{Name} is the  municipality name.
#' - \code{Total unemployment} unemployment is the  number of
   unemployed in the municipality
#' - \code{Total male unemployment} is the number of unemployed
    males in the municipality
#' - \code{Total female unemployment} is the number of
   unemployed females in the municipality
#'
#' @family Loading functions
#' @examples
#' #paro(2005,"julio","Madrid")
#'
#' @export



paro<-function(year,mes="julio",provincia){
  year<-as.character(year)
  Cap<-provincia
  provincia<-toupper(provincia)
  provincia<-a.letter(provincia)
  if(as.numeric(year)<=2011){
    if(provincia=="ARABA"){provincia1 <- "ALAVA"}
    else if(provincia=="BIZKAIA"){provincia1 <- "VIZCAYA"}
    else if(provincia=="GIPUZKOA"){provincia1 <- "GUIPUZCOA"}
    else {provincia1 <- provincia}
    } else {
      provincia1 <- provincia
      }
```

```
mes<-tolower(mes)
nn.mes<-seq(1,12,1)
names(nn.mes)<-c("enero","febrero","marzo","abril","mayo","
 junio","julio","agosto","septiembre","octubre","noviembre","
 diciembre")
cod<-paste("0",nn.mes[mes],substr(year,3,4),sep="")
dirc<-paste(getwd(),"/data_paro/",sep="")
name1<-paste(paste("MUNI",provincia1,cod,sep="_"),".xls",sep=
 "")
file1<-paste("paro_",name1,sep="")
if(sum(dir(dirc)==file1)==0){
  getbase.paro(year,mes,provincia)
}
if(as.numeric(year)<2013){
  cod1<-paste("0",nn.mes[mes],"16",sep="")
  name<-paste(paste("MUNI",provincia,cod1,sep="_"),".xls",sep
 ="")
  file<-paste("paro_",name,sep="")
  if(sum(dir(dirc)==file)==0){
    getbase.paro(2016,mes,provincia)
  }
  open<-paste(dirc,file,sep="")
  abre<-paste(dirc,file1,sep="")
  wb <- loadWorkbook(abre)
  sh <- getSheets(wb)
  hoja <- agrep("PARO",names(sh))
  datos<-xlsx::read.xlsx(abre,hoja, encoding ="UTF-8")
  idn<-xlsx::read.xlsx(open,1,colIndex=c(1:3), encoding ="UTF
 -8")
  datos<-apply(datos,2,as.character)
  idn<-apply(idn,2,as.character)
  datos<-datos[-dim(datos)[1],]
  idn<-idn[-dim(idn)[1],]
  p<-which(datos[,1]=="MUNICIPIOS")+1
  f<-dim(datos)[1]
  pi<-max(which(is.na(idn[,1])))+1
  fi<-dim(idn)[1]
  datos<-datos[p:f,]
```

```
 datos<-cbind(datos[,1],datos)
 v<-intersect(datos[,1],idn[,2])
 for( k in 1:length(v)){
   if(length(which(idn[,2]==v[k]))>1){
     datos[which(datos[,2]==v[k]),1]<-idn[which(idn[,2]==v[k
])[2],1]
   }else{
     datos[which(datos[,2]==v[k]),1]<-idn[which(idn[,2]==v[k
]),1]
   }
 }
 if(stringi::stri_count(datos[min(which(!is.na(datos[,1])))
,1],regex="[[:number:]]")==4){
   cero<-rep(0,dim(datos)[1])
   datos[,1]<-paste(cero,datos[,1],sep="")
 }
 datos<-datos[-1,]
 if(is.null(dim(datos))){datos <- t(matrix(datos ))}
 datos <- datos[,colSums(is.na(datos))!=nrow(datos)]
 if(is.null(dim(datos))){datos <- t(matrix(datos ))}
 cod<-datos[,1]
 n.m<-datos[,2]
 total<-as.numeric.factor(datos[,3])
 total.h<-as.numeric(datos[,4]) + as.numeric(datos[,5]) + as
.numeric(datos[,6])
 total.m<-as.numeric(datos[,7])+as.numeric(datos[,8])+as.
numeric(datos[,9])
 #total.h<-NA
 #total.m<-NA
 salida<-as.data.frame(cbind(cod,n.m,total,total.h,total.m))
 salida[,3:5]<-apply(salida[,3:5],2,as.numeric)
 salida[,1]<-as.character(salida[,1])
 salida[,2]<-as.character(salida[,2])
 colnames(salida)<-c("cod","nombre","paro total","paro total
 hombres","paro total mujeres")
 salida[which(is.na(salida[,3])),3]<-0
 salida[which(is.na(salida[,4])),4]<-0
 salida[which(is.na(salida[,5])),5]<-0
```

```
  #salida<-rbind(salida,rep(NA,5))
  nd<-dim(salida)[1]
  if(nd==1){
    salida <- salida
  }else{
    salida <- salida[-nd,]
  }
  #salida[nd,1]<-"Total"
  #salida[nd,2]<-Cap
  #salida[nd,3]<-sum(salida[1:nd-1,3])
  #salida[nd,4]<-sum(salida[1:nd-1,4])
  #salida[nd,5]<-sum(salida[1:nd-1,5])
  fallas<-apply(as.matrix(salida[,1]),1,nchar)
  n.fallas<-which(fallas!=5)
  if (length(n.fallas)!=0){
    for (i in seq_along(n.fallas)){
      salida[n.fallas[i],1]<-codifica(salida[n.fallas[i],2],
 provincia)
    }
  }
  salida
} else {
  abre<-paste(dirc,file1,sep="")
  wb <- loadWorkbook(abre)
  sh <- getSheets(wb)
  hoja <- agrep("PARO",names(sh))
  datos<-xlsx::read.xlsx(abre,hoja, encoding ="UTF-8")
  datos<-apply(datos,2,as.character)
  datos<-datos[-dim(datos)[1],]
  p<-max(which(is.na(datos[,1])))+1
  f<-dim(datos)[1]
  ind<-datos[p:f,1]
  if(stri_count(ind[1],regex="[[:number:]]")==4){
    cero<-rep(0,length(ind))
    ind<-paste(cero,ind,sep="")
  }
  n.m<-datos[p:f,2]
  total<-as.numeric(datos[p:f,3])
```

```
    total.h<-as.numeric(datos[p:f,4]) + as.numeric(datos[p:f
   ,5]) + as.numeric(datos[p:f,6])
    total.m<-as.numeric(datos[p:f,7])+as.numeric(datos[p:f,8])+
   as.numeric(datos[p:f,9])
    salida<-as.data.frame(cbind(ind,n.m,total,total.h,total.m))
    salida[,3:5]<-apply(salida[,3:5],2,as.numeric)
    salida[,1]<-as.character(salida[,1])
    salida[,2]<-as.character(salida[,2])
    colnames(salida)<-c("cod","nombre","paro total","paro total
    hombres","paro total mujeres")
    salida[which(is.na(salida[,3])),3]<-0
    salida[which(is.na(salida[,4])),4]<-0
    salida[which(is.na(salida[,5])),5]<-0
    #salida<-rbind(salida,rep(NA,5))
    nd<-dim(salida)[1]
    if(nd==1){
      salida <- salida
    }else{
      salida <- salida[-nd,]
    }
    #salida[nd,1]<-"Total"
    #salida[nd,2]<-Cap
    #salida[nd,3]<-sum(salida[1:nd-1,3])
    #salida[nd,4]<-sum(salida[1:nd-1,4])
    #salida[nd,5]<-sum(salida[1:nd-1,5])
    }
  salida
}
```

ALGORITHM A.23: parque.aut function code

```
#'
#' @name parque.aut
#' @rdname parque.aut
#'
#' @title Vehicle fleet information at a municipality level
#'
#' @description \code{parque.aut} collects information about
   vehicle fleet in each Spanish municipality
```

```
#'
#' @param year is numeric variable corresponding to the year,
   which must be higher than 2013
#'
#' @param ca is a character indicating one of the 17 the
   Spanish autonomous communities
#'
#' @param provincia one of the 52 Spanish provincies
#'
#' @return It is a data frame containing the following
   variables: the municipality name, number and average age of
   vehicle fleet, type of vehicles (cars, vans, trucks,
   motorcycles, buses, etc) and some other variables related to
    the register of drivers, accidents and vehicle taxes.
#'
#' @details \code{ca} may asume one of these values: "Andalucia
   ", "Asturias", "Aragon", "Baleares", "Canarias","Cantarias",
    "Castilla y Leon", "Castilla La Mancha", "Cataluña", "
   Comunidad Valenciana", "Extremadura", "Galicia", "Madrid", "
   Murcia", "Navarra", "Pais Vasco", "La Rioja", "Ceuta" and  "
   Melilla".
#'
#' @examples
#' \dontrun{parque.aut(2014,"Ceuta","Ceuta")}
#'
#' @export



parque.aut<-function(year,ca,provincia){
  year<-as.character(year)
  provincia <- toupper(provincia)
  ##Crea los directorios
  if (dir.exists(file.path(getwd(),"DGT"))==FALSE){
    dir.create(file.path(getwd(),"DGT"))
    }
  if (dir.exists(file.path(getwd(),"DGT",provincia))==FALSE){
    dir.create(file.path(getwd(),"DGT",provincia))
```

```
  }
if (dir.exists(file.path(getwd(),"DGT",provincia,year))==
 FALSE){
  dir.create(file.path(getwd(),"DGT",provincia,year))
  }
dest<-file.path(getwd(),"DGT",provincia,year)
myfiles <- NULL
# ca
ca<-tolower(ca)
# Casos raros ca
if(ca==paste("catalu","\u00F1","a",sep="")){ca<-"catalunia"}
if(ca=="comunidad valenciana"){ca<-"valencia"}
if(ca=="pais vasco"){ca<-"paisvasco"}
if(ca=="la rioja"){ca<-"la-rioja"}
if(ca=="ceuta"){ca<-"ceuta-melilla"}
if(ca=="melilla"){ca<-"ceuta-melilla"}
if(ca=="castilla y leon"){ca<-"castilla-y-leon"}
if(ca=="castilla la mancha"){ca<-"castilla-la-mancha"}

#provincia
prov<-tolower(a.letter(provincia))
# casos raros
if(prov=="araba"){prov<-"alava"}
if(prov=="ciudad_real"){prov<-"ciudad-real"}
if(prov=="a_coruna"){prov<-"corunia"}
if(prov=="las_palmas"){prov<-"las-palmas"}
if(prov=="la_rioja"){prov<-"la-rioja"}
if(prov=="tenerife"){prov <- "santa-cruz-de-tenerife"}

## carga la lista de txt
dataset<-list.files(path=dest, pattern="txt", full.names=TRUE
 )
if (length(dataset)==0){
#descarga los archivos
  page<-paste("http://www.dgt.es/es/seguridad-vial/
 estadisticas-e-indicadores/informacion-municipal/provincias/
 ",year,"/",ca,"/",prov,".shtml",sep="")
  p<-read_html(page)
```

```
 n<-xml_nodes(p,css=".tabg")
 mun<-html_attr(html_nodes(n,css="a"),"href")
 nn.mun<-str_trim(html_text(html_nodes(n,css="a")))
 aux<-sapply(mun,strsplit,"/",USE.NAMES=FALSE)
 len<-unlist(lapply(aux,length))
 files<-rep("",length(len))
 for ( i in seq_along(len)){
   files[i]<-aux[[i]][len[i]]
 }
 cod <- gsub("[^0-9]","",files)
 pat<-c(", "," ","\u00E9","\u00E1","\u00ED","\u00F3","\u00FA
","\u00C1","\u00C9","\u00CD","\u00D3","\u00DA","\u00F1","\
u00D1")
 rempl<-c("_","_","e","a","i","o","u","A","E","I","O","U","n
","N")
 for (i in 1:length(pat)){
   nn.mun<-sapply(nn.mun,str_replace_all,pat[i],rempl[i],USE
.NAMES=FALSE)
   mun<-sapply(mun,str_replace_all,pat[i],rempl[i],USE.NAMES
=FALSE)
   }
 files<-paste(year,"_",cod,"_",nn.mun,".pdf",sep="")
 for (j in 1:length(files)){
   if(sum(dir(dest)==files[j])==0){
     tryCatch({
     download.file(mun[j],paste(dest,"/",files[j],sep=""),
mode='wb')
     }, error=function(e) return(e))
   }
 }
 Sys.sleep(2)
 myfiles <- list.files(path=dest,pattern = "pdf",  full.
names = TRUE)
 lapply(myfiles, function(i) system(paste('pdftotext -eol
dos -enc UTF-8 -table', paste0('"', i, '"')), wait = FALSE)
)
 Sys.sleep(5)
 }
```

```
 dataset<-list.files(path=dest, pattern="txt", full.names=
TRUE)
 cod2 <- substr(gsub("[^0-9]","",dataset),9,nchar(dataset))
 print(length(dataset))
 base<-data.frame()
 if(as.numeric(year)<2015){
   for(k in 1:length(dataset)){
     data<-read.table(dataset[k],quote="\t",sep="\t",skip=1)
     a<-capture.output(data)
     Seg<-matrix("A",8,2)
       for (i in 4:11){
         aux<-str_trim(substring(a[i],120,nchar(a[i])))
         aux<-as.matrix(unlist(strsplit(aux," ")))
         aux<-as.vector(aux[which(nchar(aux)!=0),])
         Seg[i-3,1]<-paste0(aux[1:(length(aux)-1)],collapse=
" ")
         Seg[i-3,2]<-aux[length(aux)]
       }

     Censo.Cond<-matrix("A",3,2)
       for (i in 8:10){
         aux<-str_trim(substring(a[i],1,100))
         aux<-as.matrix(unlist(strsplit(aux," ")))
         aux<-as.vector(aux[which(nchar(aux)!=0),])
         if (i==8){
           Censo.Cond[i-7,1]<-paste0(aux[2:(length(aux)-1)],
collapse=" ")
         } else {
           Censo.Cond[i-7,1]<-paste0(c("Conductores",aux[2])
,collapse=" ")
         }
         Censo.Cond[i-7,2]<-aux[length(aux)]
       }

     sinITV<-matrix("A",3,2)
       for(i in 40:42){
         aux<-as.matrix(unlist(strsplit(a[i]," ")))
```

```
          aux<-as.vector(aux[which(nchar(aux)!=0),])
          sinITV[i-39,1]<-paste(str_trim(aux[2]),"sin ITV en
vigor",collapse=" ")
          sinITV[i-39,2]<-str_trim(aux[3])
        }

      parque<-matrix("A",6,2)
      antiguedad<-matrix("A",6,2)
        for(i in 16:21){
          aux<-as.matrix(unlist(strsplit(a[i]," ")))
          aux<-as.vector(aux[which(nchar(aux)!=0),])
          if(i==18) {aux<-aux[1:5]}
          if(i==16) {aux<-c(aux[1],paste0(aux[2:4],collapse="
 "),aux[5],"100",aux[6])}
          parque[i-15,1]<-paste0(c("Parque",aux[2]),collapse=
" ")
          antiguedad[i-15,1]<-paste0(c("Antiguedad",aux[2]),
collapse=" ")
          parque[i-15,2]<-str_trim(aux[3])
          antiguedad[i-15,2]<-str_trim(aux[5])
        }
      nn<-as.matrix(unlist(strsplit(a[2]," ")))
      nn<-as.vector(nn[which(nchar(nn)!=0),])
      nn<-nn[which(nn=="Municipio:"):(which(nn=="Provincia:")
-1)]
      if(length(nn)>2){
        Nombre<-matrix(c(nn[1],paste0(nn[2:length(nn)],
collapse=" ")),1,2)
        } else {
        Nombre<-matrix(c(nn[1:2]),1,2)
        }
      salida<-t(rbind(Nombre, parque, antiguedad, Censo.Cond,
 Seg))
      cols<-salida[1,]
      ifelse(k==1,colnames(salida)<-cols,colnames(salida)<-
names(base))
      salida<-salida[-1,]
```

```
      salida<-as.data.frame(t(salida)) ; rownames(salida)<-
NULL
      base<-rbind(base,salida)
    }
  } else {
    for(k in 1:length(dataset)){
      data<-read.table(dataset[k],quote="\t",sep="\t",skip=1)
      a<-capture.output(data)
      Seg<-matrix("A",5,2)
      for (i in 4:8){
        aux1<-str_trim(substring(a[i],120,nchar(a[i])))
        aux<-as.matrix(unlist(strsplit(aux1," ")))
        aux<-as.vector(aux[which(nchar(aux)!=0),])
        Seg[i-3,1]<-paste0(aux[1:(length(aux)-1)],collapse="
")
        Seg[i-3,2]<-aux[length(aux)]
        if(i==7){Seg[i-3,1]<-substring(aux1,str_locate(aux1,"
Sanciones")[1],str_locate(aux1,"2015")[2])}
        if(i==8){Seg[i-3,1]<-substring(aux1,str_locate(aux1,"
Puntos")[1],str_locate(aux1,"2015")[2])}
        if(Seg[i-3,2]=="2015" && nchar(aux1)<99) Seg[i-3,2]<-
"NA"
        }

    Censo.Cond<-matrix("A",3,2)
      for (i in 8:10){
        aux<-str_trim(substring(a[i],1,110))
        if(i==9) {aux<-a[i]}
        aux<-as.matrix(unlist(strsplit(aux," ")))
        aux<-as.vector(aux[which(nchar(aux)!=0),])
        Censo.Cond[i-7,1]<-paste0(c("Conductores",aux[2]),
collapse=" ")
        Censo.Cond[i-7,2]<-aux[length(aux)]
      }

    sinITV<-matrix("A",3,2)
      for(i in 38:40){
        aux<-as.matrix(unlist(strsplit(a[i]," ")))
```

```
        aux<-as.vector(aux[which(nchar(aux)!=0),])
        sinITV[i-37,1]<-paste(str_trim(aux[2]),"sin ITV en
vigor",collapse=" ")
        sinITV[i-37,2]<-str_trim(aux[3])
    }

    parque<-matrix("A",6,2)
    antiguedad<-matrix("A",6,2)
    for(i in 15:20){
      aux<-as.matrix(unlist(strsplit(a[i]," ")))
      aux<-as.vector(aux[which(nchar(aux)!=0),])
      if(i==18) {aux<-aux[1:5]}
      if(i==15) {aux<-c(aux[1],paste0(aux[2:5],collapse=" "),
aux[6],"100",aux[7])}
      parque[i-14,1]<-paste0(c("Parque",aux[2]),collapse=" ")
      antiguedad[i-14,1]<-paste0(c("Antiguedad",aux[2]),
collapse=" ")
      parque[i-14,2]<-str_trim(aux[3])
      antiguedad[i-14,2]<-str_trim(aux[5])
    }
    nn<-as.matrix(unlist(strsplit(a[2]," ")))
    nn<-as.vector(nn[which(nchar(nn)!=0),])
    nn<-nn[which(nn=="Municipio:"):(which(nn=="Provincia:")
-1)]
    if(length(nn)>2){
      Nombre<-matrix(c(nn[1],paste0(nn[2:length(nn)],collapse
=" ")),1,2)
    } else {
      Nombre<-matrix(c(nn[1:2]),1,2)
    }
    salida<-t(rbind(Nombre, parque, antiguedad, Censo.Cond,
Seg))
    cols<-salida[1,]
    ifelse(k==1,colnames(salida)<-cols,colnames(salida)<-
names(base))
    salida<-salida[-1,]
    salida<-as.data.frame(t(salida)) ; rownames(salida)<-NULL
```

```
      base<-rbind(base,salida)
      }
    }
    if(!is.null(myfiles)) {file.remove(myfiles)}
    base<-as.data.frame(cbind("Cod"=cod2,base))
  return(base)
}
```

ALGORITHM A.24: pob.a function code

```
#' @name pob.a
#' @rdname pob.a
#'
#' @title  Population grouped by age data
#' @description \code{pob.a} imports into an R file the
   municipality population database grouped by age
#'
#' @param year a numerical value from 1996 and the last
   available, which indicates the year of the required database
   .
#' @param provincia one of the 52 Spainish provinces.
#'
#' @return It is a list containing a total population data
   frame and the population grouped by sex. Each data frame
   contains the following variables::
#'
#' \itemize{
#'    \item \code{cod} is the municipality identification
   number based in the INE codification.
#'    \item \code{Name} the municipality name.
#'    \item \code{Total} total municipality population
#'    \item A hundred and one variables containing the
   population grouped by one-year age
#'  }
#'
#' @family Loading functions
#' @examples
#' pob.a(2016,"Caceres")
#'
```

```
#' @export


pob.a<-function(year,provincia){
    year<-as.character(year)
    if(as.numeric(year)<2011) stop("No existe datos para estos
   casos")
    provincia<-toupper(provincia)
    prov<-provincia
    provincia<-a.letter(provincia)
    dirc<-paste(getwd(),"/data_poblacion/",sep="")
    file<-paste(paste("pob_a",year,provincia,sep="_"),".xlsx",
   sep="")
    if(sum(dir(dirc)==file)==0){
      getbase.pob(year,provincia,anual=TRUE)
    }
    abre<-paste(dirc,file,sep="")
    datos<-xlsx::read.xlsx(abre,1, encoding ="UTF-8")
    d<-dim(datos)
    t<-which(datos[,1]=="Ambos sexos")
    h<-which(datos[,1]=="Hombres")
      if(sum(h)==0) {h<-which(datos[,1]=="Varones")}
    m<-which(datos[,1]=="Mujeres")
    edades<-datos[,2:d[2]]
    nn<-edades[which(edades[,1]=="Total"),]
    nn<-apply(nn,1,as.character)
    colnames(edades)<-nn
    nombres<-as.character(datos[,1])
    codigo<-rep("AA",d[1])
    municipio<-rep("AA",d[1])
    for (i in 1:d[1]){
      nn<-unlist(strsplit(nombres[i],"-"))
      codigo[i]<-str_trim(nn[1])
      municipio[i]<-str_trim(nn[2])
    }
    salida<-cbind(codigo,municipio,edades)
    salida[,1]<-as.character(salida[,1])
    salida[,2]<-as.character(salida[,2])
```

```
    salida[,3:dim(salida)[2]]<-apply(salida[,3:dim(salida)
   [2]],2,as.numeric.factor)
    dd<-h-t-1
    s.t<-salida[(t+1):(h-1),]
    s.h<-salida[(h+1):(m-1),]
    s.m<-salida[(m+1):(m+dd),]
    out<-list(s.t[-1,],s.h[-1,],s.m[-1,])
    names(out)<-c("Ambos Sexos","Hombre","Mujeres")
    out
}
```

ALGORITHM A.25: pob.e.ev function code

```
#' @name pob.e.ev
#' @rdname pob.e.ev
#'
#' @title Panel of foreign population at a municipality level
   for a period of time
#'
#' @description \code{pob.e.ev} creates a data frame with a
   panel of  total foreign population at a municipality level
   for a period of time from the  \code{inicio} to \code{fin}.
#' @param inicio starting year of the panel, which must be
   higher than 1996.
#' @param fin last year of the panel.
#' @param provincia one of the 52 Spanish provinces.
#' @param print logical variable  do  you need print a output
   file with the results?   for which FALSE is the default
   value
#'
#' @return It is a data frame
#'
#' @details If \code{print} is set to \code{TRUE}, an \code{
   xlsx} file containing the data frame is saved int the folder
    \code{Outputs} called \code{pob_foreign_ev_provincia_inicio
   -fin.xlsx}
#'
#' @family Manipulate functions
#' @examples
```

```
#' pob.ev(2005,2007,"Avila")
#'
#' @export



pob.e.ev<-function(inicio,fin,provincia,print=FALSE){
  if(fin<inicio) stop("La fecha de inicio debe ser mayor que la
    fecha de fin")
  n<-seq(inicio,fin,1)
  year<-as.character(sort(n,decreasing=TRUE))
  base<-pob.e.tot(year[1],provincia)
  for (i in 2:length(year)){
    aux<-rep(NA,dim(base)[1])
    pob<-pob.e.tot(year[i],provincia)
    v<-intersect(base[,1],pob[,1])
    for(j in 1:length(v)){
      aux[which(base[,1]==v[j])]<-pob[which(pob[,1]==v[j]),3]
    }
    base<-cbind(base,aux)
    }
  colnames(base)<-c("Cod","Municipio",year)
  orden<-c(1,2,seq(dim(base)[2],3))
  base<-base[,orden]
    if (print==TRUE){
      if(dir.exists(file.path(getwd(),"Outputs"))==FALSE){
      dir.create(file.path(getwd(),"Outputs"))
      }
    file<-paste(getwd(),"/Outputs/pob_foreign_ev_",provincia,"_
   ",paste(inicio,fin,sep="-"),".xlsx",sep="")
    xlsx::write.xlsx(base,file)
    }
  base
}
```

Algorithm A.26: pob.e function code

```
#' @name pob.e
#' @rdname pob.e
```

```
#'
#' @title  Population grouped by nationality and sex
#' @description \code{pob.e} imports into R the municipality
   population grouped by nationality and sexa
#'
#' @param year a numerical value from 1996 and the latest
   available, which indicates the year of the required database
   .
#' @param provincia one of the 52 Spainish provinces.
#'
#' @return It is a list containing a data frame with total
   population and population by sex. Each data frame contains
   the following variables:
#' \itemize{
#'    \item \code{cod} is the municipality identification
   number based in the INE codification.
#'    \item \code{Name} is the municipality name.
#'    \item \code{Total} is the municipality population.
#'    \item Three variables containing information about
   population by age (younger than 16 years old, from 16 to 64
   years old and older than 65 years old), which are, in turn,
   grouped by total population, foreign population and national
    population.
#'  }
#'
#' @family Loading functions
#' @examples
#' pob.e(2016,"Madrid")
#'
#' @export

pob.e<-function(year,provincia){
    year<-as.character(year)
    provincia<-toupper(provincia)
    prov<-provincia
    provincia<-a.letter(provincia)
    dirc<-paste(getwd(),"/data_poblacion/",sep="")
```

```
  file<-paste(paste("pob_e",year,provincia,sep="_"),".xlsx",
sep="")
  if(sum(dir(dirc)==file)==0){
    getbase.pob(year,provincia,extr=TRUE)
  }
  abre<-paste(dirc,file,sep="")
  datos<-xlsx::read.xlsx(abre,1, encoding ="UTF-8")
  d<-dim(datos)
  t<-which(datos[,1]=="Ambos sexos")
  h<-which(datos[,1]=="Hombres")
    if(sum(h)==0) {h<-which(datos[,1]=="Varones")}
  m<-which(datos[,1]=="Mujeres")
  edades<-datos[,2:d[2]]
  nombres<-as.character(datos[,1])
  codigo<-rep("AA",d[1])
  municipio<-rep("AA",d[1])
  for (i in 1:d[1]){
    nn<-unlist(strsplit(nombres[i],"-"))
    codigo[i]<-nn[1]
    municipio[i]<-nn[2]
  }
  salida<-cbind(codigo,municipio,edades)
  salida[,1]<-as.character(salida[,1])
  salida[,2]<-as.character(salida[,2])
  salida[,3:dim(salida)[2]]<-apply(salida[,3:dim(salida)
[2]],2,as.numeric.factor)
  fila<-c("Cod","Municipio","Total","Total Menores de 16 a\
u00F1os","Total De 16 a 64 a\u00F1os","Total De 65 y mas a\
u00F1os", "Total Esp","Esp Menores de 16 a\u00F1os","Esp De
16 a 64 a\u00F1os","Esp De 65 y mas a\u00F1os","Total Extr",
"Extr Menores de 16 a\u00F1os","Extr De 16 a 64 a\u00F1os","
Extr De 65 y mas a\u00F1os")
  colnames(salida)<-fila
  dd<-h-t-1
  s.t<-salida[(t+1):(h-1),]
  s.h<-salida[(h+1):(m-1),]
  s.m<-salida[(m+1):(m+dd),]
  out<-list(s.t[-1,],s.h[-1,],s.m[-1.])
```

```
    names(out)<-c("Ambos Sexos","Hombre","Mujeres")
    out
}
```

ALGORITHM A.27: pob.e.tot function code

```
#' @name pob.e.tot
#' @rdname pob.e.tot
#'
#' @title  Total Foreign Population
#' @description \code{pob.e.tot} imports into an R file the
   municipality total foreign population.
#'
#' @param year a numerical value from 1996 and the latest
   available year, which indicates the year of the required
   database.
#' @param provincia one of the 52 Spainish provinces.
#'
#' @return It is a data frame containing the municipality total
    foreign population
#'
#' @family Loading functions
#' @examples
#' pob.e.tot(2016,"Madrid")
#'
#' @export

pob.e.tot<-function(year,provincia){
    year<-as.character(year)
    provincia<-toupper(provincia)
    prov<-provincia
    provincia<-a.letter(provincia)
    dirc<-paste(getwd(),"/data_poblacion/",sep="")
    file<-paste(paste("pob_e",year,provincia,sep="_"),".xlsx",
   sep="")
    if(sum(dir(dirc)==file)==0){
      getbase.pob(year,provincia,extr=TRUE)
    }
    abre<-paste(dirc,file,sep="")
```

```
 datos<-xlsx::read.xlsx(abre,1,colIndex=c(1,4,5,10,11),
encoding ="UTF-8")
 d<-dim(datos)
 nombres<-as.character(datos[,1])
 codigo<-rep("AA",d[1])
 municipio<-rep("AA",d[1])
 if (as.numeric(year)<2006){
   for (i in 1:d[1]){
   nn<-unlist(strsplit(nombres[i]," "))
   codigo[i]<-str_trim(nn[5])
   municipio[i]<-str_trim(nn[6])
   }
 } else {
   for (i in 1:d[1]){
   nn<-unlist(strsplit(nombres[i],"-"))
   codigo[i]<-str_trim(nn[1])
   municipio[i]<-str_trim(nn[2])
   }
 }
   t<-which(datos[,1]=="Ambos sexos")
   h<-which(datos[,1]=="Hombres")
     if(sum(h)==0) {h<-which(datos[,1]=="Varones")}
   edades<-datos[,2:d[2]]
   salida<-cbind(codigo,municipio,edades)
   salida[,1]<-as.character(salida[,1])
   salida[,2]<-as.character(salida[,2])
   salida[,3:dim(salida)[2]]<-apply(salida[,3:dim(salida)
[2]],2,as.numeric.factor)
   salida[which(is.na(salida[,3])),3]<-0
   salida[which(is.na(salida[,4])),4]<-0
   salida[which(is.na(salida[,5])),5]<-0
   salida[which(is.na(salida[,6])),6]<-0
   fila<-c("Cod","Municipio","Total")
   colnames(salida)<-fila
   if(as.numeric(year)<2002){
     s.t<-salida[t:(h-1),1:3]
   } else {
     s.t<-salida[t:(h-1),c(1,2,5)]
```

```
      }
    s.t<-s.t[-c(1:2),]
    #s.t[1,2]<-"Todos"
    s.t
}
```

ALGORITHM A.28: pob.ev function code

```
#' @name pob.ev
#' @rdname pob.ev
#'
#' @title Panel of total population at the municipality level
   for a period of time
#'
#' @description Create a data frame containing a panel of total
    population at the municipality level from the years \code{
   inicio} to \code{fin}.
#' @param inicio starting year of the panel, which must be
   higher than 1996.
#' @param fin last year of the panel.
#' @param provincia one of the 52 Spanish provinces.
#' @param print logical variable  do  you need to print a
   output file with the results?   being FALSE is the default
   value.
#'
#' @return It is a data frame
#'
#' @details If \code{print} is set to  \code{TRUE}, an \code{
   xlsx} file containing the data frame is saved in the folder
   \code{Outputs} called \code{pob_total_ev_provincia_inicio-
   fin.xlsx}
#'
#' @family Manipulate functions
#' @examples
#' pob.ev(2005,2007,"Avila")
#'
#' @export

pob.ev<-function(inicio,fin,provincia,print=FALSE){
```

```
  if(fin<inicio) stop("La fecha de inicio debe ser mayor que la
    fecha de fin")
  n<-seq(inicio,fin,1)
  year<-as.character(sort(n,decreasing=TRUE))
  base<-pob.tot(year[1],provincia)
  for (i in 2:length(year)){
    aux<-rep(NA,dim(base)[1])
    pob<-pob.tot(year[i],provincia)
    v<-intersect(base[,1],pob[,1])
    for(j in 1:length(v)){
      aux[which(base[,1]==v[j])]<-pob[which(pob[,1]==v[j]),3]
    }
    base<-cbind(base,aux)
    }
  colnames(base)<-c("Cod","Municipio",year)
  orden<-c(1,2,seq(dim(base)[2],3))
  base<-base[,orden]
    if (print==TRUE){
      if(dir.exists(file.path(getwd(),"Outputs"))==FALSE){
      dir.create(file.path(getwd(),"Outputs"))
      }
    file<-paste(getwd(),"/Outputs/pob_total_ev_",provincia,"_",
   paste(inicio,fin,sep="-"),".xlsx",sep="")
    xlsx::write.xlsx(base,file)
    }
  base
}
```

Algorithm A.29: pob.fen.ev function code

```
#' @name pob.fen.ev
#' @rdname pob.fen.ev
#'
#' @title Panel of the number of births and deaths at a
   municipality level for a period of time
#'
```

```
#' @description \code{pob.fen.ev} creates a list with the panel
    of the number of births and deaths of Spain at the
   municipality level for a period of time from the years  \
   code{inicio} to \code{fin}.
#' @param inicio starting year of the panel, which must be
   higher than 1996..
#' @param fin last year of the panel.
#' @param provincia one of the 52 Spanish provinces.
#' @param print logical variable  do  you need print a output
   file with the results?   for which FALSE is the default
   value.
#'
#' @return It is list containing two data frames for the number
    of births  (Nacimientos) and deaths (Falleciemientos) at
   the municpality level.
#'
#' @details If \code{print} is set to \code{TRUE}, an \code{
   xlsx} file containing two sheet, one per each variable data
   frame, is saved into the folder \code{Outputs} called \code{
   fen_evol_total_provincia_inicio-fin.xlsx}
#'
#' @family Manipulate functions
#' @examples
#' pob.fen.ev(2005,2007,"Avila")
#'
#' @export


pob.fen.ev<-function(inicio,fin,provincia,print=FALSE){
  if(fin<inicio) stop("La fecha de inicio debe ser mayor que la
    fecha de fin")
  n<-seq(inicio,fin,1)
  year<-as.character(sort(n,decreasing=TRUE))
  base<-pob.fen(year[1],provincia)
  base.n<-cbind(base[,c(1,2,3)])
  base.f<-cbind(base[,c(1,2,4)])
  for (i in 2:length(year)){
    aux1<-rep(NA,dim(base)[1])
```

```
    aux2<-rep(NA,dim(base)[1])
    pob<-pob.fen(year[i],provincia)
    v<-intersect(base[,1],pob[,1])
    for(j in 1:length(v)){
      aux1[which(base[,1]==v[j])]<-pob[which(pob[,1]==v[j]),3]
      aux2[which(base[,1]==v[j])]<-pob[which(pob[,1]==v[j]),4]
    }
    base.n<-cbind(base.n,aux1)
    base.f<-cbind(base.f,aux2)
    }
  colnames(base.n)<-c("Cod","Municipio",year)
  colnames(base.f)<-c("Cod","Municipio",year)
  orden<-c(1,2,seq(dim(base.n)[2],3))
  base.n<-base.n[,orden]
  base.f<-base.f[,orden]
  if(print==TRUE){
      if(dir.exists(file.path(getwd(),"Outputs"))==FALSE){
      dir.create(file.path(getwd(),"Outputs"))
      }
    excel<-createWorkbook()
    s1<-createSheet(excel,sheetName="Nacimientos")
    s2<-createSheet(excel,sheetName="Fallacimientos")
    addDataFrame(base.n,s1)
    addDataFrame(base.f,s2)
    file<-paste(getwd(),"/Outputs/fen_evol_total_",provincia,"_
   ",paste(inicio,fin,sep="-"),".xlsx",sep="")
    saveWorkbook(excel,file)
  }
  base<-list(base.n,base.f)
  names(base)<-c("Nacimientos","Fallecimientos")
  base
}
```

Algorithm A.30: pob.fen function code

```
#' @importFrom stats complete.cases
#' @name pob.fen
#' @rdname pob.fen
#'
```

```
#' @title  Number of births and deaths at a municipality level
#' @description \code{pob.fen} imports into an R file two main
   demographic indexes at the municipality level: number of
   births and deaths.
#'
#' @param year a numerical value from 1996 and 2015, which
   indicates the year of the required database
#' @param provincia one of the 52 Spainish province.
#'
#' @return It is a data frame containing two main demographic
   indexes:
#'  \itemize{
#'    \item \code{Birth} number of birth in the municipality
#'    \item \code{Deaths} Number of death in the municipality
#'  }
#'
#' @family Loading functions
#' @examples
#' pob.fen(2012,"Madrid")
#'
#' @export


pob.fen<-function(year,provincia){
    year<-as.character(year)
    provincia<-toupper(provincia)
    prov<-provincia
    provincia<-a.letter(provincia)
    dirc<-paste(getwd(),"/data_poblacion/",sep="")
    file<-paste(paste("fen",year,provincia,sep="_"),".xlsx",sep
   ="")
    if(sum(dir(dirc)==file)==0){
      getbase.fen(year,provincia)
    }
    abre<-paste(dirc,file,sep="")
    datos<-xlsx::read.xlsx(abre,1,colIndex=c(1,2,5), encoding =
   "UTF-8")
    datos<-datos[which(complete.cases(datos)==TRUE),]
```

```
    datos<-datos[-1,]
    d<-dim(datos)
    nombres<-as.character(datos[,1])
    codigo<-rep("AA",d[1])
    municipio<-rep("AA",d[1])
    for (i in 1:d[1]){
      nn<-unlist(strsplit(nombres[i]," "))
      codigo[i]<-str_trim(nn[1])
      if(length(nn)>2){
        nom<-paste0(nn[2:length(nn)],collapse=" ")
        municipio[i]<-nom
      } else {
      municipio[i]<-str_trim(nn[2])
      }
    }
    cifras<-as.data.frame(datos[,2:3])
    cifras<-apply(cifras,2,as.numeric.factor)
    if (is.null(dim(cifras))){ cifras <- t(as.matrix(cifras)) }
    ids<-as.data.frame(cbind(codigo,municipio))
    salida<-cbind(ids,cifras)
    colnames(salida)<-c("Cod","Municipio","Nacidos","Fallecidos
   ")
    salida
}
```

Algorithm A.31: pob.h.ev function code

```
#' @name pob.h.ev
#' @rdname pob.h.ev
#'
#' @title Panel of the male population at the municipality
   level for a period of time
#'
#' @description pob.h.ev. creates a data frame containing the
   panel of the male population of Spain at the municipality
   level for a period of time from the years \code{inicio} to \
   code{fin}.
#' @param inicio starting year of the panel, which must be
   higher than 1996.
```

```
#' @param fin last year of the panel.
#' @param provincia one of the 52 Spanish provinces.
#' @param print logical variable  do  you need to print output
    file with the results? being FALSE the default value.
#'
#' @return It is a data frame
#'
#' @details If \code{print} is set to \code{TRUE}, an \code{
   xlsx} file containing the data frame is saved in the folder
   \code{Outputs} called \code{pob_male_ev_provincia_inicio-fin
   .xlsx}
#'
#' @family Manipulate functions
#' @examples
#' pob.h.ev(2005,2007,"Avila")
#'
#' @export


pob.h.ev<-function(inicio,fin,provincia,print=FALSE){
  if(fin<inicio) stop("La fecha de inicio debe ser mayor que la
    fecha de fin")
  n<-seq(inicio,fin,1)
  year<-as.character(sort(n,decreasing=TRUE))
  base<-pob.h.tot(year[1],provincia)
  for (i in 2:length(year)){
    aux<-rep(NA,dim(base)[1])
    pob<-pob.h.tot(year[i],provincia)
    v<-intersect(base[,1],pob[,1])
    for(j in 1:length(v)){
      aux[which(base[,1]==v[j])]<-pob[which(pob[,1]==v[j]),3]
    }
    base<-cbind(base,aux)
    }
  colnames(base)<-c("Cod","Municipio",year)
  orden<-c(1,2,seq(dim(base)[2],3))
  base<-base[,orden]
    if (print==TRUE){
```

```
      if(dir.exists(file.path(getwd(),"Outputs"))==FALSE){
      dir.create(file.path(getwd(),"Outputs"))
      }
    file<-paste(getwd(),"/Outputs/pob_male_ev_",provincia,"_",
   paste(inicio,fin,sep="-"),".xlsx",sep="")
    xlsx::write.xlsx(base,file)
    }
  base
}
```

Algorithm A.32: pob.h.tot function code

```
#' @name pob.h.tot
#' @rdname pob.h.tot
#'
#' @title  Male Population  data
#' @description \code{pob.h.tot} imports into an R file the
   male population of Spain by municipality.
#'
#' @param year a numerical value from 1996 and the latest
   available year, which indicates the year of the required
   database.
#' @param provincia one of the 52 Spainish provinces.
#'
#' @return It is a data frame containing the municipality male
   population
#'
#' @family Loading functions
#' @examples
#' pob.h.tot(2016,"Madrid")
#'
#' @export


pob.h.tot<-function(year,provincia){
    year<-as.character(year)
    provincia<-toupper(provincia)
    prov<-provincia
    provincia<-a.letter(provincia)
```

```
  dirc<-paste(getwd(),"/data_poblacion/",sep="")
  file<-paste(paste("pob_q",year,provincia,sep="_"),".xlsx",
sep="")
  if(sum(dir(dirc)==file)==0){
    getbase.pob(year,provincia)
  }
  abre<-paste(dirc,file,sep="")
  datos<-xlsx::read.xlsx(abre,1,colIndex=c(1:4), encoding ="
UTF-8")
  d<-dim(datos)
  nombres<-as.character(datos[,1])
  codigo<-rep("AA",d[1])
  municipio<-rep("AA",d[1])
  if (as.numeric(year)<2006){
    for (i in 1:d[1]){
    nn<-unlist(strsplit(nombres[i]," "))
    codigo[i]<-str_trim(nn[5])
    municipio[i]<-str_trim(nn[6])
    }
  } else {
    for (i in 1:d[1]){
    nn<-unlist(strsplit(nombres[i],"-"))
    codigo[i]<-str_trim(nn[1])
    municipio[i]<-str_trim(nn[2])
    }
  }
  t<-which(datos[,1]=="Ambos sexos")
  h<-which(datos[,1]=="Hombres")
    if(sum(h)==0) {h<-which(datos[,1]=="Varones")}
  m<-which(datos[,1]=="Mujeres")
  edades<-datos[,2:d[2]]
  salida<-cbind(codigo,municipio,edades)
  salida[,1]<-as.character(salida[,1])
  salida[,2]<-as.character(salida[,2])
  salida[,3:4]<-apply(salida[,3:4],2,as.numeric.factor)
  fila<-c("Cod","Municipio","Total")
  colnames(salida)<-fila
  s.h<-salida[h:(m-1),1:3]
```

```
    s.h<-s.h[-c(1:2),]
    #s.h[1,2]<-"Todos"
    s.h
}
```

ALGORITHM A.33: pob.ind.p function code

```
#' @name pob.ind.p
#' @rdname pob.ind.p
#'
#' @title Calculate some demographic indexes
#' @description \code{pob.ind.p} computes a set of demographic
   indexes
#' @param year a numerical value from 1996 and the latest
   available year, which indicates the year of the required
   database.
#' @param provincia one of the 52 Spanish provinces.
#' @param print logical variable  do  you need to print a
   output file with the results?   being FALSE the default
   value.
#'
#' @return A data frame contains the spatial units, in rows,
   and the demographic index computed in  columns.
#'
#' @details This function calculates ten demographical indexes:
#' \itemize{
#'  \item Childhood index
#'  \item Youth index
#'  \item Third age index
#'  \item Dependence index
#'  \item Unemployment rate, both sexes
#'  \item Unemployment rate, males
#'  \item Unemployment rate, females
#'  \item Municipality average age, both sexes
#'  \item Municipality average age, males
#'  \item Municipality average age, females
#' }
#' For a full description of the index, see the 2017
   Socioeconomic Atlas of Extremadura
```

```
#'
#' If \code{print} is set to \code{TRUE}, an \code{xlsx} file
   containing ten sheet -one  data frame per index- is saved
   into the folder \code{Outputs} which the name  \code{pob_
   index_provincia_year.xlsx}
#'The principal difference between this function and \link{pob.
   ind} is the precision of the calculation of the municipality
    average age. Since the year 2011, this function uses the
   one-year age variable for the computation instead of the
   five-year age approximation used in \code{pob.ind}. For more
    details, see the 2017 Socioeconomic Atlas of Extremadura.
#'
#' @references{
#' Junta de Extremadura (2017) Atlas Socieconómico de
   Extremadura 2017. Mérida (Spain).
#' \url{http://estadistica.gobex.es/web/guest/atlas-
   socieconomico-de-extremadura}
#'}
#' @family Manipulate functions
#' @examples
#' pob.ind.p(2012,"Madrid")
#'
#' @export



pob.ind.p<-function(year,provincia,print=FALSE){
  if(year<2011) stop("No existe datos para estos casos")
  input<-pob.a(year,provincia)
  base<-input[[1]]
  base.h<-input[[2]]
  base.m<-input[[3]]
  t<-dim(base)
  infancia<-(apply(base[,4:18],1,sum)/base[,3])*100
  vejez<-(apply(base[,69:dim(base)[2]],1,sum)/base[3])*100
  juventud<-(apply(base[,19:33],1,sum)/base[,3])*100
  dependencia<-(apply(base[,c(4:19,69:t[2])],1,sum)/apply(base
   [,20:68],1,sum))*100
```

```
edad<-c(0.5,seq(1,101,1))
par<-rep(NA,t[1])
par.h<-rep(NA,t[1])
par.m<-rep(NA,t[1])
if(year>2005){
  d<-paro(year,provincia=provincia)
  v<-intersect(base[,1],d[,1])
  for(j in 1:length(v)){
    b<-which(base[,1]==v[j])
    e<-which(d[,1]==v[j])
    par[b]<-(d[e,3]/apply(base[b,20:68],1,sum))*100
    par.h[b]<-(d[e,4]/apply(base.h[b,20:68],1,sum))*100
    par.m[b]<-(d[e,5]/apply(base.m[b,20:68],1,sum))*100
    }
  }
media<-rep(NA,t[1])
media.h<-rep(NA,t[1])
media.m<-rep(NA,t[1])
for(k in 1:t[1]){
  media[k]<-sum(base[k,4:t[2]]*edad)/base[k,3]
  media.h[k]<-sum(base.h[k,4:t[2]]*edad)/base.h[k,3]
  media.m[k]<-sum(base.m[k,4:t[2]]*edad)/base.m[k,3]
}
salida<-cbind(base[,1:2],infancia,juventud,vejez,dependencia,
 par,par.h,par.m,media,media.h,media.m)
colnames(salida)<-c("Cod","Municipio","Infancia","Juventud","
 Vejez","Dependencia","Paro Total","Paro Hombres","Paro
 Mujeres","Edad Media","Edad Media Homres","Edad Media
 Mujeres")
num<-sapply(salida,is.numeric)
salida[num]<-apply(salida[num],2,round,1)
if(print==TRUE){
    if(dir.exists(file.path(getwd(),"Outputs"))==FALSE){
    dir.create(file.path(getwd(),"Outputs"))
    }
file<-paste(getwd(),"/Outputs/pob_index-p_",provincia,"_",
 year,".xlsx",sep="")
write.xlsx(salida,file)
```

```
    }
  salida
}
```

ALGORITHM A.34: pob.ind function code

```
#' @name pob.ind
#' @rdname pob.ind
#'
#' @title Calculate the population indexs
#'
#' @description \code{pob.ind} computes a set of demographic
   indexes
#' @param year a numerical value from 1996 and the latest
   available year, which indicates the year of the required
   database.
#' @param provincia one of the 52 Spanish provinces.
#' @param print logical variable  do  you need to print a
   output file with the results?   being FALSE the default
   value.
#'
#' @return A data frame contains the spatial units, in rows,
   and the demographic index computed in  columns.
#'
#' @details This function calculates ten demographical indexes:
#' \itemize{
#'  \item Childhood index
#'  \item Youth index
#'  \item Third age index
#'  \item Dependence index
#'  \item Unemployment rate, both sexes
#'  \item Unemployment rate, males
#'  \item Unemployment rate, females
#'  \item Municipality average age, both sexes
#'  \item Municipality average age, males
#'  \item Municipality average age, females
#' }
#' For a full description of the index, see the 2017
   Socioeconomic Atlas of Extremadura
```

```
#'
#' If \code{print} is set to \code{TRUE}, an \code{xlsx} file
   containing ten sheet -one  data frame per index- is saved
   into the folder \code{Outputs} which the name  \code{pob_
   index_provincia_year.xlsx}
#' @references{
#' Junta de Extremadura (2017) Atlas Socieconómico de
   Extremadura 2017. Mérida (Spain).
#' \url{http://estadistica.gobex.es/web/guest/atlas-
   socieconomico-de-extremadura}
#'}
#' @family Manipulate functions
#' @examples
#' pob.ind(2012,"Madrid")
#'
#' @export


pob.ind<-function(year,provincia,print=FALSE){
  if(year>2011){cat("Puede obtener un indice mas preciso usando
    la funcion pob.ind.p() \n")}
  entrada<-pob.q(year,provincia)
  base<-entrada[[1]]
  base.h<-entrada[[2]]
  base.m<-entrada[[3]]
  t<-dim(base)
  infancia<-(apply(base[,4:6],1,sum)/base[,3])*100
  vejez<-(apply(base[,17:t[2]],1,sum)/base[3])*100
  juventud<-(apply(base[,7:9],1,sum)/base[,3])*100
  dependencia<-(apply(base[,c(4:6,17:t[2])],1,sum)/apply(base
   [,7:16],1,sum))*100
  par<-rep(NA,t[1])
  par.h<-rep(NA,t[1])
  par.m<-rep(NA,t[1])
  if(year>2005){
    d<-paro(year,provincia=provincia)
    v<-intersect(base[,1],d[,1])
    for(j in 1:length(v)){
```

```
      b<-which(base[,1]==v[j])
      e<-which(d[,1]==v[j])
      par[b]<-(d[e,3]/apply(base[b,7:16],1,sum))*100
      par.h[b]<-(d[e,4]/apply(base.h[b,7:16],1,sum))*100
      par.m[b]<-(d[e,5]/apply(base.m[b,7:16],1,sum))*100
      }
    }
  edad<-c(seq(2,97,5),101)
  media<-rep(NA,t[1])
  media.h<-rep(NA,t[1])
  media.m<-rep(NA,t[1])
  for(k in 1:t[1]){
    media[k]<-sum(base[k,4:t[2]]*edad)/base[k,3]
    media.h[k]<-sum(base.h[k,4:t[2]]*edad)/base.h[k,3]
    media.m[k]<-sum(base.m[k,4:t[2]]*edad)/base.m[k,3]
  }
  salida<-cbind(base[,1:2],infancia,juventud,vejez,dependencia,
   par,par.h,par.m,media,media.h,media.m)
  colnames(salida)<-c("Cod","Municipio","Infancia","Juventud","
   Vejez","Dependencia","Paro Total","Paro Hombres","Paro
   Mujeres","Edad Media","Edad Media Homres","Edad Media
   Mujeres")
  num<-sapply(salida,is.numeric)
  salida[num]<-apply(salida[num],2,round,1)
  if(print==TRUE){
      if(dir.exists(file.path(getwd(),"Outputs"))==FALSE){
      dir.create(file.path(getwd(),"Outputs"))
      }
  file<-paste(getwd(),"/Outputs/pob_index_",provincia,"_",year,
   ".xlsx",sep="")
  write.xlsx(salida,file)
  }
  salida
}
```

Algorithm A.35: pob.m.ev function code

```
#' @name pob.m.ev
#' @rdname pob.m.ev
```

```
#'
#' @title Panel of the female population at the municipality
   level for a period of time
#'
#' @description pov.m.ev create a data frame containing the
   panel of the female population of Spain at the municipality
   level for a period of time from the  \code{inicio} to \code{
   fin}.
#' @param inicio starting year of the panel, which must be
   higher than 1996..
#' @param fin last year of the panel.
#' @param provincia one of the 52 Spanish provinces.
#' @param print logical variable  do  you need to print a
   output file with the results?   being FALSE the default
   value.
#'
#' @return It is a data frame
#'
#' @details If \code{print} is set to \code{TRUE}, a \code{xlsx
   } file containing the data frame is saved in the folder \
   code{Outputs} which name \code{pob_female_ev_provincia_
   inicio-fin.xlsx}
#'
#' @family Manipulate functions
#' @examples
#' pob.m.ev(2005,2007,"Avila")
#'
#' @export



pob.m.ev<-function(inicio,fin,provincia,print=FALSE){
  if(fin<inicio) stop("La fecha de inicio debe ser mayor que la
    fecha de fin")
  n<-seq(inicio,fin,1)
  year<-as.character(sort(n,decreasing=TRUE))
  base<-pob.m.tot(year[1],provincia)
  for (i in 2:length(year)){
```

```
    aux<-rep(NA,dim(base)[1])
    pob<-pob.m.tot(year[i],provincia)
    v<-intersect(base[,1],pob[,1])
    for(j in 1:length(v)){
      aux[which(base[,1]==v[j])]<-pob[which(pob[,1]==v[j]),3]
    }
    base<-cbind(base,aux)
    }
  colnames(base)<-c("Cod","Municipio",year)
  orden<-c(1,2,seq(dim(base)[2],3))
  base<-base[,orden]
    if (print==TRUE){
      if(dir.exists(file.path(getwd(),"Outputs"))==FALSE){
      dir.create(file.path(getwd(),"Outputs"))
      }
    file<-paste(getwd(),"/Outputs/pob_female_ev_",provincia,"_"
   ,paste(inicio,fin,sep="-"),".xlsx",sep="")
    xlsx::write.xlsx(base,file)
    }
  base
}
```

ALGORITHM A.36: pob.m.tot function code

```
#' @name pob.m.tot
#' @rdname pob.m.tot
#'
#' @title  Female Population  data
#' @description \code{pob.m.tot} imports into an R file the
   female population of Spain by municipality.
#'
#' @param year a numerical value from 1996 and the latest
   available year, which indicates the year of the required
   database.
#' @param provincia one of the 52 Spainish provinces.
#'
#' @return  It is a  data frame conatining the municipality
   total female population
#'
```

```
#' @family Loading functions
#' @examples
#' pob.m.tot(2016,"Madrid")
#'
#' @export



pob.m.tot<-function(year,provincia){
    year<-as.character(year)
    provincia<-toupper(provincia)
    prov<-provincia
    provincia<-a.letter(provincia)
    dirc<-paste(getwd(),"/data_poblacion/",sep="")
    file<-paste(paste("pob_q",year,provincia,sep="_"),".xlsx",
  sep="")
    if(sum(dir(dirc)==file)==0){
      getbase.pob(year,provincia)
    }
    abre<-paste(dirc,file,sep="")
    datos<-xlsx::read.xlsx(abre,1,colIndex=c(1:4), encoding ="
  UTF-8")
    d<-dim(datos)
    nombres<-as.character(datos[,1])
    codigo<-rep("AA",d[1])
    municipio<-rep("AA",d[1])
    if (as.numeric(year)<2006){
      for (i in 1:d[1]){
      nn<-unlist(strsplit(nombres[i]," "))
      codigo[i]<-str_trim(nn[5])
      municipio[i]<-str_trim(nn[6])
      }
    } else {
      for (i in 1:d[1]){
      nn<-unlist(strsplit(nombres[i],"-"))
      codigo[i]<-str_trim(nn[1])
      municipio[i]<-str_trim(nn[2])
```

```
      }
    }
    t<-which(datos[,1]=="Ambos sexos")
    h<-which(datos[,1]=="Hombres")
      if(sum(h)==0) {h<-which(datos[,1]=="Varones")}
    m<-which(datos[,1]=="Mujeres")
    edades<-datos[,2:d[2]]
    salida<-cbind(codigo,municipio,edades)
    salida[,1]<-as.character(salida[,1])
    salida[,2]<-as.character(salida[,2])
    salida[,3:4]<-apply(salida[,3:4],2,as.numeric.factor)
    fila<-c("Cod","Municipio","Total")
    colnames(salida)<-fila
    dd<-h-t-1
    s.m<-salida[m:(m+dd),1:3]
    s.m<-s.m[-c(1:2),]
    #s.m[1,2]<-"Todos"
    s.m
}
```

ALGORITHM A.37: pob.n.ev function code

```
#' @name pob.n.ev
#' @rdname pob.n.ev
#'
#' @title Panel of the national population at the municipality
   level for a period of time
#'
#' @description \code{pob.n.ev} creates a data frame containing
    the panel of the national population of Spain at the
   municipality level for a period of time from the years  \
   code{inicio} to \code{fin}.
#' @param inicio starting year of the panel, which must be
   higher than 1996
#' @param fin last year of the panel.
#' @param provincia one of the 52 Spanish provinces.
#' @param print logical variable  do  you need to print a
   output file with the results?   being FALSE the default
   value
```

```
#'
#' @return It is a data frame
#'
#' @family Loading functions
#' @details If \code{print} is set to \code{TRUE}, an \code{
   xlsx} file containing the data frame is saved in the folder
   \code{Outputs} which name \code{pob_national_ev_provincia_
   inicio-fin.xlsx}
#'
#' @family Manipulate functions
#' @examples
#' pob.n.ev(2005,2007,"Avila")
#'
#' @export



pob.n.ev<-function(inicio,fin,provincia,print=FALSE){
  if(fin<inicio) stop("La fecha de inicio debe ser mayor que la
    fecha de fin")
  n<-seq(inicio,fin,1)
  year<-as.character(sort(n,decreasing=TRUE))
  base<-pob.n.tot(year[1],provincia)
  for (i in 2:length(year)){
    aux<-rep(NA,dim(base)[1])
    pob<-pob.n.tot(year[i],provincia)
    v<-intersect(base[,1],pob[,1])
    for(j in 1:length(v)){
      aux[which(base[,1]==v[j])]<-pob[which(pob[,1]==v[j]),3]
    }
    base<-cbind(base,aux)
    }
  colnames(base)<-c("Cod","Municipio",year)
  orden<-c(1,2,seq(dim(base)[2],3))
  base<-base[,orden]
    if (print==TRUE){
      if(dir.exists(file.path(getwd(),"Outputs"))==FALSE){
      dir.create(file.path(getwd(),"Outputs"))
```

```
      }
    file<-paste(getwd(),"/Outputs/pob_national_ev_",provincia,"
   _",paste(inicio,fin,sep="-"),".xlsx",sep="")
    xlsx::write.xlsx(base,file)
    }
  base
}
```

ALGORITHM A.38: pob.n.tot function code

```
#' @name pob.n.tot
#' @rdname pob.n.tot
#'
#' @title  National Population  data
#' @description \code{pob.n.tot} imports into an R file the
   national population of Spain by municipality.
#'
#' @param year a numerical value from 1996 and the latest
   available year, which indicates the year of the required
   database.
#' @param provincia one of the 52 Spainish provinces.
#'
#' @return It is a data frame containing the municipality
   national population of Spain.
#'
#' @family Loading functions
#' @examples
#' pob.n.tot(2016,"Madrid")
#'
#' @export


pob.n.tot<-function(year,provincia){
    year<-as.character(year)
    provincia<-toupper(provincia)
    prov<-provincia
    provincia<-a.letter(provincia)
    dirc<-paste(getwd(),"/data_poblacion/",sep="")
```

```
 file<-paste(paste("pob_e",year,provincia,sep="_"),".xlsx",
sep="")
 if(sum(dir(dirc)==file)==0){
   getbase.pob(year,provincia,extr=TRUE)
 }
 abre<-paste(dirc,file,sep="")
 datos<-xlsx::read.xlsx(abre,1,colIndex=c(1,2,6,7), encoding
="UTF-8")
 d<-dim(datos)
 nombres<-as.character(datos[,1])
 codigo<-rep("AA",d[1])
 municipio<-rep("AA",d[1])
 if (as.numeric(year)<2006){
   for (i in 1:d[1]){
   nn<-unlist(strsplit(nombres[i]," "))
   codigo[i]<-str_trim(nn[5])
   municipio[i]<-str_trim(nn[6])
   }
 } else {
   for (i in 1:d[1]){
   nn<-unlist(strsplit(nombres[i],"-"))
   codigo[i]<-str_trim(nn[1])
   municipio[i]<-str_trim(nn[2])
   }
 }
 t<-which(datos[,1]=="Ambos sexos")
 h<-which(datos[,1]=="Hombres")
   if(sum(h)==0) {h<-which(datos[,1]=="Varones")}
 edades<-datos[,2:d[2]]
 salida<-cbind(codigo,municipio,edades)
 salida[,1]<-as.character(salida[,1])
 salida[,2]<-as.character(salida[,2])
 salida[,3:5]<-apply(salida[,3:5],2,as.numeric.factor)
 fila<-c("Cod","Municipio","Total")
 colnames(salida)<-fila
 if (as.numeric(year)<2002){
   s.t<-salida[t:(h-1),1:3]
 } else {
```

```
      s.t<-salida[t:(h-1),c(1,2,4)]
    }
    s.t<-s.t[-c(1:2),]
    #s.t[1,2]<-"Todos"
    s.t

}
```

ALGORITHM A.39: pob.q function code

```
#' @importFrom stringr str_trim
#' @name pob.q
#' @rdname pob.q
#'
#' @title Population by five-year age groups and sex
#' @description \code{pob.q} imports into an R file the Spanish
    municipality population by five-year age groups and sex
#'
#' @param year a numerical value from 1996 and the latest
   available year, which indicates the year of the required
   database.
#' @param provincia one of the 52 Spainish provinces.
#'
#' @return It is a list containing a data frame of the Spanish
   municipality population by five-year age groups and sex.
   Each data frame contains the following variables::
#' - \code{cod} is the municipality identification number based
    in the INE codification.
#'  - \code{Name} is the municipality name.
#' - \code{Total} is the municipality population.
#' - Twenty-one variables containing the population by five-
   year age groups.
#'
#' @family Loading functions
#' @examples
#' pob.q(2016,"Caceres")
#'
#' @export
```

```
pob.q<-function(year,provincia){
    year<-as.character(year)
    provincia<-toupper(provincia)
    prov<-provincia
    provincia<-a.letter(provincia)
    dirc<-paste(getwd(),"/data_poblacion/",sep="")
    file<-paste(paste("pob_q",year,provincia,sep="_"),".xlsx",
   sep="")
    if(sum(dir(dirc)==file)==0){
      getbase.pob(year,provincia)
    }
    abre<-paste(dirc,file,sep="")
    datos<-xlsx::read.xlsx(abre,1, encoding ="UTF-8")
    d<-dim(datos)
    t<-which(datos[,1]=="Ambos sexos")
    h<-which(datos[,1]=="Hombres")
      if(sum(h)==0) {h<-which(datos[,1]=="Varones")}
    m<-which(datos[,1]=="Mujeres")
    edades<-datos[,2:d[2]]
    nn<-edades[which(edades[,1]=="Total"),]
    nn<-apply(nn,1,as.character)
    colnames(edades)<-nn
    nombres<-as.character(datos[,1])
    codigo<-rep("AA",d[1])
    municipio<-rep("AA",d[1])
    if (as.numeric(year)<2006){
      for (i in 1:d[1]){
      nn<-unlist(strsplit(nombres[i]," "))
      codigo[i]<-str_trim(nn[5])
      municipio[i]<-str_trim(nn[6])
      }
    } else {
      for (i in 1:d[1]){
      nn<-unlist(strsplit(nombres[i],"-"))
      codigo[i]<-str_trim(nn[1])
      municipio[i]<-str_trim(nn[2])
      }
    }
```

```
    salida<-cbind(codigo,municipio,edades)
    salida[,1]<-as.character(salida[,1])
    salida[,2]<-as.character(salida[,2])
    salida[,3:dim(salida)[2]]<-apply(salida[,3:dim(salida)
   [2]],2,as.numeric.factor)
    dd<-h-t-1
    s.t<-salida[(t+1):(h-1),]
    s.h<-salida[(h+1):(m-1),]
    s.m<-salida[(m+1):(m+dd),]
    out<-list(s.t[-1,],s.h[-1,],s.m[-1,])
    names(out)<-c("Ambos Sexos","Hombre","Mujeres")
    out
}
```

Algorithm A.40: pob.tot function code

```
#' @name pob.tot
#' @rdname pob.tot
#'
#' @title Population  data
#' @description \code{pob.tot} imports into an R file the total
    population of Spain by municipality.
#'
#' @param year a numerical value from 1996 and the latest
   available year, which indicates the year of the required
   database
#' @param provincia one of the 52 Spainish provinces.
#'
#' @return It is a data frame containing the total population
   of Spain at the municipality level.
#'
#' @family Loading functions
#' @examples
#' pob.tot(2016,"Madrid")
#'
#' @export

pob.tot<-function(year,provincia){
    year<-as.character(year)
```

```
  provincia<-toupper(provincia)
  prov<-provincia
  provincia<-a.letter(provincia)
  dirc<-paste(getwd(),"/data_poblacion/",sep="")
  file<-paste(paste("pob_q",year,provincia,sep="_"),".xlsx",
sep="")
  if(sum(dir(dirc)==file)==0){
    getbase.pob(year,provincia)
  }
  abre<-paste(dirc,file,sep="")
  datos<-xlsx::read.xlsx(abre,1,colIndex=c(1:4), encoding ="
UTF-8")
  d<-dim(datos)
  nombres<-as.character(datos[,1])
  codigo<-rep("AA",d[1])
  municipio<-rep("AA",d[1])
  if (as.numeric(year)<2006){
    for (i in 1:d[1]){
    nn<-unlist(strsplit(nombres[i]," "))
    codigo[i]<-str_trim(nn[5])
    municipio[i]<-str_trim(nn[6])
    }
  } else {
    for (i in 1:d[1]){
    nn<-unlist(strsplit(nombres[i],"-"))
    codigo[i]<-str_trim(nn[1])
    municipio[i]<-str_trim(nn[2])
    }
  }
  t<-which(datos[,1]=="Ambos sexos")
  h<-which(datos[,1]=="Hombres")
    if(sum(h)==0) {h<-which(datos[,1]=="Varones")}
  m<-which(datos[,1]=="Mujeres")
  edades<-datos[,2:d[2]]
  salida<-cbind(codigo,municipio,edades)
  salida[,1]<-as.character(salida[,1])
  salida[,2]<-as.character(salida[,2])
  salida[,3:4]<-apply(salida[,3:4],2,as.numeric.factor)
```

```
    fila<-c("Cod","Municipio","Total")
    colnames(salida)<-fila
    s.t<-salida[t:(h-1),1:3]
    s.t<-s.t[-c(1:2),]
    #s.t[1,2]<-"Todos"
    s.t
}
```

ALGORITHM A.41: simpleCap function code

```
# Coloca la primer letra en mayuscula de un vector de
   caracteres


simpleCap <- function(x) {
    s <- strsplit(x, " ")[[1]]
    paste(substring(s, 1, 1), tolower(substring(s, 2)),
          sep = "", collapse = " ")
}
```

B Reproducible source code of "The dynamics of patentability and collaborativeness in Chile: an analysis of social networks between 1989 and 2013"

Algorithm B.1: Final Paper source code

```
#______packages________#

library(openxlsx) #allow the use os xlsx in R
library(sna)
library(network)
library(GGally)
library(gridExtra)

#__Loading data_________#

db0 <- read.xlsx("clean_BBDD.xlsx")
b<-read.xlsx("/Users/andresvallone/Dropbox/Patents/Data/Codigos
    applicants.xlsx",1)
b<-rbind(b, c("A.M.I.E. (CHILE) LTDA.",2929))
b[2079,1]<-"PUCCIO SALVADOR ANTONIO"

################################################
#                Network                       #
#        take at least 45 min to finish        #
################################################

#full network

net<-table(db0$id_pat,db0$app_name)
```

```
net.bin <- ifelse(net>0,1,0)

#Coding assignees

cod<-rep(0,dim(db0)[1])
aux<-as.character(b[,1])
for (i in 1:length(cod)){
  cod[i]<-match(db0$app_name[i],aux)
}

aux1<-colnames(net)
id<-b[match(aux1,aux),2]
nn<-b[match(id,b[,2]),1]

colnames(net)<-as.character(id)


######    Warning this part delete assignee from the network
     #######
#   1904 patent and applicant was deleted from the net.bin
#   for beeing a non cooperation patent and beeing the only
#   patent of the applicant
#
   ##########################################################################


tot<-rowSums(net.bin)
net.1<-cbind(net.bin,tot)
net.1<-rbind(net.1,colSums(net.1))

for(j in 1:(ncol(net.1)-1)){
  print(j)
  for(i in 1:(nrow(net.1)-1)){
    if (net.1[i,j]==1 && net.1[nrow(net.1),j]==1 && net.1[i,
   ncol(net.1)]==1) net.1[i,j]<-0 else net.1[i,j]<-net.1[i,j]
  }
}
```

```
net.1<-net.1[-nrow(net.1),]
net.1<-net.1[,!colnames(net.1)=="tot"]

tot<-rowSums(net.1)
net.1<-cbind(net.1,tot)

net.f<-subset(net.1,net.1[,ncol(net.1)]>0)
net.f<-net.f[,!colnames(net.f)=="tot"]

net.f<-t(net.f)
tot<-rowSums(net.f)
net.f<-cbind(net.f,tot)
net.f<-subset(net.f,net.f[,ncol(net.f)]>0)
net.f<-net.f[,!colnames(net.f)=="tot"]
net.f<-t(net.f)

###### Plots #######
# Bipartite network.

redtotal<-network(t(net.f),bipartite=TRUE, matrix.type="
   bipartite")
tt<-colSums(net.f)
l.a<-rep("",length(tt))
l.e<-rep("",dim(net.f)[1])
l.a[which(tt>60)]<-sub("X","",names(tt)[which(tt>60)])
t1<-sub("X","",names(tt)[which(tt>60)])
t2<-rep("",length(t1))
for (i in 1:length(t1)){
  t2[i]<-b[which(b[,2]==t1[i]),1]
}

tabla<-as.data.frame(cbind(t1,t2))
colnames(tabla)<-c("ID","Applicant")
row.names(tabla)<-NULL

mytheme <- ttheme_minimal(
    core = list(fg_params=list(cex = .8,hjust=0,x=0)),
    colhead = list(fg_params=list(cex = .8,hjust=0,x=0)),
```

```
    rowhead = list(fg_params=list(col="white",cex =.8)))

tc<-tableGrob(tabla,theme=mytheme)

col = c("actor" = "grey30", "event" = "black")
shp = c("actor" = 15, "event" = 19)
g1<-ggnet2(redtotal,node.size=.5,color="mode",edge.col="grey70"
   ,palette=col,shape="mode",label=c(l.a,l.e),legend.position="
   none",shape.palette=shp,label.color="black",label.size=5,
   node.alpha=.2)


png("fig7.png",width =800, height = 600)
grid.arrange(tc,g1,ncol=2,widths=c(.5,1))
dev.off()


#Co-affiliation netwrok.

adj <- tcrossprod(t(net.f))
f1<-rowSums(adj-diag(adj)*diag(dim(adj)[1]))!=0
adj1<-adj[f1,f1]

c.a.net <- network(adj1,directed=FALSE)

ggnet2(c.a.net, layout.par = list(cell.jitter = 0.5),node.size
   =3,color="grey50",shape=15,edge.col="black",label=c(sub('X',
   '',colnames(adj1))),vjust=-1,label.size=3,label.alpha=0.75)

#Selection cluster

coop.net <- adj1

ids<-rownames(coop.net)

u1<-c(which(coop.net[which(ids=="X2642"),]>0))
u2<-c(which(coop.net[which(ids=="X593"),]>0),which(ids=="X2860"
   )); names(u2)[length(u2)]<-"X2860"
```

```
u3<-c(which(coop.net[which(ids=="X1750"),]>0))
u4<-c(which(coop.net[which(ids=="X2919"),]>0),which(ids=="X1228
   ")); names(u4)[length(u4)]<-"X1228"
u5<-c(which(coop.net[which(ids=="X41"),]>0)) #bloque cerrado
u6<-c(which(coop.net[which(ids=="X1717"),]>0))
u7<-c(which(coop.net[which(ids=="X1893"),]>0),which(ids=="X1796
   "));names(u7)[length(u7)]<-"X1796"
u8<-c(which(coop.net[which(ids=="X2868"),]>0))
u9<-c(which(coop.net[which(ids=="X286"),]>0))
u10<-c(which(coop.net[which(ids=="X61"),]>0))
u11<-c(which(coop.net[which(ids=="X2895"),]>0))
u12<-c(which(coop.net[which(ids=="X2918"),]>0))
u13<-c(which(coop.net[which(ids=="X2922"),]>0))
u14<-c(83,73,7,115); names(u14)<-ids[u14]
u15<-c(79,210,88,219); names(u15)<-ids[u15]

  pal<-c("chocolate","cadetblue","cornflowerblue","blue4","
   burlywood3","darkgreen","hotpink4","darkred","deeppink3","
   peachpuff3","orangered2","khaki4","navajowhite4","violetred2
   ","green1")


col<-rep("grey50",231)
col[u1]<-pal[1]
col[u2]<-pal[2]
col[u3]<-pal[3]
col[u4]<-pal[4]
col[u5]<-pal[5]
col[u6]<-pal[6]
col[u7]<-pal[7]
col[u8]<-pal[8]
col[u9]<-pal[9]
col[u10]<-pal[10]
col[u11]<-pal[11]
col[u12]<-pal[12]
col[u13]<-pal[13]
col[u14]<-pal[14]
col[u15]<-pal[15]
```

```
c1<-network(coop.net[u1,u1],names.eval="pat",directed=FALSE)
e1<-coop.net[u1,u1]
diag(e1)<-0
times<-e1[which(e1>0)]
set.edge.value(c1,"times",times[1:(length(times)/2)])
c2<-network(coop.net[u2,u2],names.eval="pat",directed=FALSE)
e1<-coop.net[u2,u2]
diag(e1)<-0
times<-e1[which(e1>0)]
set.edge.value(c2,"times",times[1:(length(times)/2)])
c3<-network(coop.net[u3,u3],names.eval="pat",directed=FALSE)
e1<-coop.net[u3,u3]
diag(e1)<-0
times<-e1[which(e1>0)]
set.edge.value(c3,"times",times[1:(length(times)/2)])
c4<-network(coop.net[u4,u4],names.eval="pat",directed=FALSE)
e1<-coop.net[u4,u4]
diag(e1)<-0
times<-e1[which(e1>0)]
set.edge.value(c4,"times",times[1:(length(times)/2)])
c5<-network(coop.net[u5,u5],names.eval="pat")
e1<-coop.net[u5,u5]
diag(e1)<-0
set.edge.value(c5,"times",e1[which(e1>0)])
c6<-network(coop.net[u6,u6],names.eval="pat")
e1<-coop.net[u6,u6]
diag(e1)<-0
set.edge.value(c6,"times",e1[which(e1>0)])
c7<-network(coop.net[u7,u7],names.eval="pat")
e1<-coop.net[u7,u7]
diag(e1)<-0
set.edge.value(c7,"times",e1[which(e1>0)])
c8<-network(coop.net[u8,u8],names.eval="pat")
e1<-coop.net[u8,u8]
diag(e1)<-0
set.edge.value(c8,"times",e1[which(e1>0)])
c9<-network(coop.net[u9,u9],names.eval="pat")
```

```
e1<-coop.net[u9,u9]
diag(e1)<-0
set.edge.value(c9,"times",e1[which(e1>0)])
c10<-network(coop.net[u10,u10],names.eval="pat")
e1<-coop.net[u10,u10]
diag(e1)<-0
set.edge.value(c10,"times",e1[which(e1>0)])
c11<-network(coop.net[u11,u11],names.eval="pat")
e1<-coop.net[u11,u11]
diag(e1)<-0
set.edge.value(c11,"times",e1[which(e1>0)])
c12<-network(coop.net[u12,u12],names.eval="pat")
e1<-coop.net[u12,u12]
diag(e1)<-0
set.edge.value(c12,"times",e1[which(e1>0)])
c13<-network(coop.net[u13,u13],names.eval="pat")
e1<-coop.net[u13,u13]
diag(e1)<-0
set.edge.value(c13,"times",e1[which(e1>0)])
c14<-network(coop.net[u14,u14],names.eval="pat")
e1<-coop.net[u14,u14]
diag(e1)<-0
set.edge.value(c14,"times",e1[which(e1>0)])
c15<-network(coop.net[u15,u15],names.eval="pat")
e1<-coop.net[u15,u15]
diag(e1)<-0
set.edge.value(c15,"times",e1[which(e1>0)])


coop.net_1<-network(coop.net,names.eval="pat")
e1<-coop.net
diag(e1)<-0
d<-e1[which(e1>0)]
d[which(d==6)]<-4
d[which(d==24)]<-5
set.edge.value(coop.net_1,"times",+.5*d)

#graph co-affiliation final.
```

```
ggnet2(coop.net_1,mode = "fruchtermanreingold", layout.par =
   list(cell.jitter = 0.5),node.size=3,node.col=col,shape=15,
   edge.col="black",label=c(sub('X','',colnames(cola))),vjust
   =-1,label.size=3,label.alpha=0.75)

#extracted network graph


l1<-b[match(sub("X","",names(u1)),b[,2]),1]
l2<-b[match(sub("X","",names(u1)),b[,2]),2]
l2[3]<-l1[3]
l2[13]<-l1[13]
l2[5]<-l1[5]
cl1<-ggnet2(c1,node.size=4,node.col=pal[1],shape=15,edge.col="
   black",label=l2,vjust=2,label.size=5,label.alpha=0.95,layout
   .exp=0.3,edge.label="times",edge.label.size=4)

cl2<-ggnet2(c2,node.size=4,node.col=pal[2],shape=15,edge.col="
   black",label=b[match(sub("X","",names(u2)),b[,2]),1],vjust
   =2,label.size=5,label.alpha=0.95,layout.exp=0.3,,edge.label=
   "times",edge.label.size=4)

cl3<-ggnet2(c3,node.size=4,node.col=pal[3],shape=15,edge.col="
   black",label=b[match(sub("X","",names(u3)),b[,2]),1],vjust
   =2,label.size=5,label.alpha=0.95,layout.exp=0.4,edge.label="
   times",edge.label.size=4)

cl3<-ggnet2(c3,node.size=4,node.col=pal[3],shape=15,edge.col="
   black",label=b[match(sub("X","",names(u3)),b[,2]),1],vjust
   =2,label.size=4,label.alpha=0.95,layout.exp=0.4,edge.label=k
   ,edge.label.size=4)


l41<-b[match(sub("X","",names(u4)),b[,2]),1]
l42<-b[match(sub("X","",names(u4)),b[,2]),2]
l42[6]<-l41[6]
```

```
cl4<-ggnet2(c4,node.size=4,node.col=pal[4],shape=15,edge.col="
   black",label=l42,vjust=2,label.size=5,label.alpha=0.95,
   layout.exp=0.3,,edge.label="times",edge.label.size=4)


cl6<-ggnet2(c6,node.size=4,node.col=pal[6],shape=15,edge.col="
   black",label=b[match(sub("X","",names(u6)),b[,2]),1],vjust
   =2,label.size=3,label.alpha=0.95,layout.exp=0.3,,edge.label=
   "times",edge.label.size=4)

cl7<-ggnet2(c7,node.size=4,node.col=pal[7],shape=15,edge.col="
   black",label=b[match(sub("X","",names(u7)),b[,2]),1],vjust
   =2,label.size=3,label.alpha=0.95,layout.exp=0.3,,edge.label=
   "times",edge.label.size=4)

cl5<-ggnet2(c5,node.size=4,node.col=pal[5],shape=15,edge.col="
   black",label=b[match(sub("X","",names(u5)),b[,2]),1],vjust
   =2,label.size=3,label.alpha=0.95,layout.exp=0.3,edge.label="
   times",edge.label.size=4)


cl8<-ggnet2(c8,node.size=4,node.col=pal[8],shape=15,edge.col="
   black",label=b[match(sub("X","",names(u8)),b[,2]),1],vjust
   =2,label.size=3,label.alpha=0.95,layout.exp=0.3,edge.label="
   times", edge.alpha=.35)

cl9<-ggnet2(c9,node.size=4,node.col=pal[9],shape=15,edge.col="
   black",label=b[match(sub("X","",names(u9)),b[,2]),1],vjust
   =2,label.size=3,label.alpha=0.95,layout.exp=0.3,edge.label="
   times", edge.alpha=.35)

cl10<-ggnet2(c10,node.size=4,node.col=pal[10],shape=15,edge.col
   ="black",label=b[match(sub("X","",names(u10)),b[,2]),1],
   vjust=2,label.size=3,label.alpha=0.95,layout.exp=0.3,edge.
   label="times", edge.alpha=.35)
```

```
cl11<-ggnet2(c11,node.size=4,node.col=pal[11],shape=15,edge.col
   ="black",label=b[match(sub("X","",names(u11)),b[,2]),1],
   vjust=2,label.size=3,label.alpha=0.95,layout.exp=0.3,edge.
   label="times", edge.alpha=.35)

cl12<-ggnet2(c12,node.size=4,node.col=pal[12],shape=15,edge.col
   ="black",label=b[match(sub("X","",names(u12)),b[,2]),1],
   vjust=2,label.size=3,label.alpha=0.95,layout.exp=0.3,edge.
   label="times", edge.alpha=.35)

cl13<-ggnet2(c13,node.size=4,node.col=pal[13],shape=15,edge.col
   ="black",label=b[match(sub("X","",names(u13)),b[,2]),1],
   vjust=2,label.size=3,label.alpha=0.95,layout.exp=0.3,edge.
   label="times", edge.alpha=.35)

cl14<-ggnet2(c14,node.size=4,node.col=pal[14],shape=15,edge.col
   ="black",label=b[match(sub("X","",names(u14)),b[,2]),1],
   vjust=2,label.size=3,label.alpha=0.95,layout.exp=0.3,edge.
   label="times", edge.alpha=.35)

cl15<-ggnet2(c15,node.size=4,node.col=pal[15],shape=15,edge.col
   ="black",label=b[match(sub("X","",names(u15)),b[,2]),1],
   vjust=2,label.size=3,label.alpha=0.95,layout.exp=0.3,edge.
   label="times", edge.alpha=.35)


grid.arrange(cl1,cl2,cl3,cl4,ncol=2,nrow=2,padding = unit( 2, "
   line"))



png("fig9d.png",width =800, height = 600)
cl4
dev.off()
```

```
#Prod vs. colla spatial analysis.

w <- ifelse(adj1>0,1,0)
diag(w)<-0

library(spdep)

W <- mat2listw(w)
B <- mat2listw(w,style="B")
prod <- diag(adj1)

l.m <- localmoran(prod,W,alternative="two.sided")
MI <- moran.mc(prod,W,9999)
G <- globalG.test(prod,B)

quad <- function(x,W,y,geoda=TRUE){
  if(missing(y)){y <- x}
  if(geoda==TRUE){
    qn <- c("hh"=1,"ll"=2,"hl"=4,"lh"=3)
  } else {
    qn <- c("hh"=1,"ll"=3,"hl"=4,"lh"=2)
  }
  zx <- (x - mean(x))#/sd(x)
  lx <- spdep::lag.listw(W,y)
  zlx <- (lx - mean(lx))#/sd(lx)
  hx <- zx > 0
  hlx <-  zlx > 0
  hh <- hx * hlx
  lh <- (1 - hx) * hlx
  hl <- hx * (1 - hlx)
  ll <- (1 - hx) * (1 - hlx)
  out <- qn["hh"]*hh + qn["ll"]*ll + qn["hl"]*hl + qn["lh"]*lh
  return(out)
}


l.m.1<-quad(prod,W)
```

```
col.1 <- l.m.1
col.1[col.1==1]<-"HH"
col.1[col.1=="2"]<-"LL"
col.1[col.1=="3"]<-"LH"
col.1[col.1=="4"]<-"HL"

c.a.net %v% "quad"=col.1

ggnet2(c.a.net,mode = "fruchtermanreingold", layout.par = list(
   cell.jitter = 0.5),node.size=3,color="quad",shape=15,edge.
   col="black",label=c(sub('X','',colnames(adj1))),vjust=-1,
   label.size=3,label.alpha=0.75,palette=c("HH"="red","LL"="
   blue","LH"="pink","HL"="lightblue"),size.cut=2)
```

C “msp” package source code

Algorithm C.1: msp function code

```
# @import stringi
#' @import stringdist
#' @name msp
#' @rdname msp

#'
#' @title Fix misspelling and different in similar character
   data.
#'

#' @param x a character to search
#' @param y the vector containing x and order character
#' @param r level of similarity between element, by default is
   0.8
#' @param type an character indicating the replacement
   procedure. Se details for more information
#' @param ... additional arguments are passed on to \link{
   stringsim} function.
#' @description Look for similarities between a single
   character entry and a character vector to detect
   similarities dues to misspeling problems and fix it.
#'
#' @return a same object than \code{x} without misspelling
#' @details Type define the way of the rigth option to replace
   is choose, |code{ask} generate a interactive replacement
   process, \code{first} use the fisrt element of the list as
   correct element to repalce and \code{min} use the smallest
   element as correct to replace. The default value is \code{
   ask}
```

```
#' @examples
#' \dontrun{
#'  set.seed(1)
#'  z <- stringi::stri_rand_strings(20, 5)
#'  a <- z[1]
#'  a1 <- c("GNZuC","GNZuC ", "GNZu.C"," GNZuC", "GNZ uC", "
   gNZuC")
#'  z <- c(z,sample(z),sample(z,10),a1)
#'  msp(a,z)
#' }
#' @export

msp <- function(x,y,r=0.8,type = "ask",...){
    change <- NULL
    goodone <- NULL
    add <- NULL
    skips <- NULL
    #checks
    if(length(a)>1) stop("Please provide a single element to
   search")
    if(is.character(x)!=TRUE) stop(" You must provide a
   character")
    if(!is.character(z)) stop("The target vector must be a
   character vector")
    #type definitions
    type.ask<-function(search){
      display<-c(search, "None")
      cat(display, fill=max(nchar(search)) ,labels= paste0("[",
   seq_along(display),"]"))
      goodone <<- readline("Please select the correct option to
    replace: ")
      if ((length(search)+1)==as.numeric(goodone)){
        add <<- readline("Please introduce the correct option
   to replace: ")
        goodone <<- length(search)
      }
      skip <- readline("Do you want to skip any opcion? Yes (Y)
    No (N): ")
```

```
    if (skip =="Y"){
      skips <- readline("Please indicate the registers to
skip: ")
      skips <<- as.integer(unlist(strsplit(skips," ")))
    }
  }

  type.first<-function(search){
    goodone <<- 1L
  }

  type.min<-function(search){
    long <- sapply(search,nchar)
    goodone <<- which(long==min(long))[1L]
  }

  #global function
  clean.y <- y
  unique.y <- unique(y)
  #search <- agrep(x,unique.y, max.distance = r,value=TRUE)
  sim <- stringdist::stringsim(x,unique.y,...)
  target <- which(sim >= r)
  search <- unique.y[target]
  change <- seq_along(search)
  if(length(search)==0) stop("No mactching, nothing to
replace")
  if (length(search)==1) {
     type.first(search)
  } else if (type =="ask"){
     type.ask(search)
  } else if (type=="first"){
     type.first(search)
  } else  {
     type.min(search)
  }
  if(!is.null(add)){
    replace<-add
  } else {
```

```
      replace <- search[as.numeric(goodone)]
    }
    for ( j in change){
      if(j %in% skips) next
      clean.y[which(y==search[change[j]])] <- replace
    }
  return(clean.y)
}
```

D Reproducible source code of "Spatio-temporal methods for the analysis of the Chilean urban system dynamics"

Algorithm D.1: Final paper source code

```
#Hay que ajustar los path hasta /PAPERS/

setwd("/Users/andresvallone/Dropbox/tesis_AndresVallone/PAPERS/
   Paper 1_ZipfLaw/Data/SHP/c5000")

require(xlsx)
require(spdep)
require(Kendall)
require(maptools)
require(rgdal)
library(ggplot2)
library(devtools)
library(reshape2)
library(RColorBrewer)
library(ggmap)
library(gridExtra)

den<-function(x){ #log calc and mean normalization
  y<-log(x)
  z<-mean(y)
  out<-y/z
  out
}
```

```
Year<-c(1930,1940,1952,1960,1970,1982,1992,2002)
set.seed(2010)

base<-readOGR(getwd(),"c5000")

base$P70[180] <- 1525 #fix data problems in La islita


bRM<-subset(base,base$FIELD2=="RM")
bnoRM<-subset(base,!base$FIELD2=="RM")


b <- as.data.frame(base)
d.bRM <- as.data.frame(bRM)
d.bnoRM <- as.data.frame(bnoRM)
b <- as.data.frame(b[,6:13])
b <- apply(b,2,as.numeric)
colnames(b)<-Year
d.bRM <- d.bRM[,6:13]
d.bnoRM <- d.bnoRM[,6:13]
d.noStgo <- d.bRM[-1,]


#fig 1

coords<-coordinates(base)
pt <- as.data.frame(coords)

lon <- c(-77,-65)
lat <- c(-55,-18)

bbox <- make_bbox(lon,lat,f=3)
map.chile1 <- get_map(bbox,maptype = "terrain")
```

```
m1 <- ggmap(map.chile1)+geom_point(data=pt,aes(x=coords.x1,y=
   coords.x2),col="red",shape=21,size=1)+scale_x_continuous(
   limits = lon, expand = c(0, 0),breaks=NULL) + scale_y_
   continuous(limits = lat, expand = c(0, 0),breaks=NULL)+xlab(
   "")+ylab("")


d.e <- function(x){
  t <- summary(x)
  nt <- names(t)
  se <- sd(x)
  out <- round(c(t[nt],"s.d"=se),2)
  out
}


d.table <- apply(b,2,d.e)


pop  <- as.vector(as.matrix(b))
yr <- unlist(lapply(seq_along(Year),function(x) rep(Year[x
   ],184)))
data <- data.frame("Year"=as.character(yr),"Population"=pop)


labs <- unlist(lapply(seq_along(Year),function(x) paste(Year[x
   ],"\n","\n",paste(d.table[,x],collapse="\n"),collapse="\n"))
   )

# asegura que le numeros de caracteres del vector de 1930 sea
   el mismo para todos elementos.

n1 <- nchar(d.table[,1])
d2 <- rownames(d.table)
t <- c(0,4,4,6,5,4,3)
for(k in seq_along(d2)){
  e <- n1[k]+t[k]
```

```
  d2[k] <- stringr::str_pad(d2[k],width=(26-e),side="right",pad
   =" ")
}
y1<-stringr::str_pad(Year[1],width=27,side="left",pad=" ")
y1 <- paste(y1,"\n","\n",sep="")
d1 <- paste(d2,d.table[,1],collapse="\n")
l1 <- c(y1,d1)
l1 <- paste(l1,collapse="")

labs[1] <- l1

 b.plot <- ggplot(data,aes(x=Year,y=Population,group=Year))+
   geom_point(col="blue",shape=1,size=2)+stat_summary(fun.y=
   mean, geom="point", shape=21, size=4,col="red",fill="green")
   +stat_summary(geom = "errorbar",fun.data = mean_cl_normal)+
   scale_x_discrete(name="",labels=labs)+theme_classic()+ theme
   (axis.line = element_line(linetype="blank"),axis.text.x=
   element_text(hjust=c(0.95,rep(0.9,7))))

quartz()
grid.arrange(m1,b.plot, ncol = 2,widths=c(.5,1))



# fig 2

MR <- colSums(d.bRM)
Rest <-colSums(d.bnoRM)
Country <- colMeans(rbind(MR,Rest))
wMR <- colSums(d.noStgo)
dat.f1.a <- rbind(Country,MR,Rest)
colnames(dat.f1.a) <- Year

dat.f2.a<-data.frame(dat.f1.a, group=c("Country Avarage","MR","
   Rest"))
dat.f2.a <- melt(dat.f1.a,id.var=c("group"))
colnames(dat.f1.a) <- c("group","Year","Population")

quartz()
```

```
fig2.a <- ggplot(dat.f1.a,aes(x=Year,y=Population,group=group,
   colour=group))+geom_line()+geom_point()
fig2.a


dat.f1.b <- rbind(Country,Rest,wMR)
colnames(dat.f1.b) <- Year
dat.f1.b <- data.frame(dat.f1.b,group=c("Country Avarage","All
   the Rest","MR without Santiago"))
dat.f1.b <- melt(dat.f1.b,id.var=c("group"))
colnames(dat.f1.b) <- c("group","Year","Population")

quartz()
fig1.b <- ggplot(dat.f1.b,aes(x=Year,y=Population,group=group,
   colour=group))+geom_line()+geom_point()

# fig 3

D <- apply(b,2,den)
DRM <- apply(d.bRM,2,den)
DnoRM <- apply(d.bnoRM,2,den)

D<-as.data.frame(D)
names(D)<-Year

DRM<-as.data.frame(DRM)
names(DRM)<-Year

DnoRM<-as.data.frame(DnoRM)
names(DnoRM)<-Year


#graph parameters
m.X<-rep(0,ncol(D))
m.Y<-rep(0,ncol(D))
M.X<-rep(0,ncol(D))
M.Y<-rep(0,ncol(D))
```

```
m.X.RM<-rep(0,ncol(D))
m.Y.RM<-rep(0,ncol(D))
M.X.RM<-rep(0,ncol(D))
M.Y.RM<-rep(0,ncol(D))


m.X.noRM<-rep(0,ncol(D))
m.Y.noRM<-rep(0,ncol(D))
M.X.noRM<-rep(0,ncol(D))
M.Y.noRM<-rep(0,ncol(D))

# Tamaño de los ejes

for(i in 1:8){
  d<-density(D[,i])
  m.X[i]<-min(d$x)
  m.Y[i]<-min(d$y)
  M.X[i]<-max(d$x)
  M.Y[i]<-max(d$y)
}


for(i in 1:8){
  d<-density(DnoRM[,i])
  m.X.noRM[i]<-min(d$x)
  m.Y.noRM[i]<-min(d$y)
  M.X.noRM[i]<-max(d$x)
  M.Y.noRM[i]<-max(d$y)
}


for(i in 1:8){
  d<-density(DRM[,i])
  m.X.RM[i]<-min(d$x)
  m.Y.RM[i]<-min(d$y)
  M.X.RM[i]<-max(d$x)
  M.Y.RM[i]<-max(d$y)
```

```
}

a.x<-c(min(m.X),max(M.X))
a.y<-c(min(m.Y),max(M.Y))
#202
col<-c("red","springgreen4","blue") #colores

quartz()
plot(density(DnoRM[,1]),xlim=c(0.4,max(M.X.noRM)),ylim=c(min(m.
   Y.noRM),max(M.Y.noRM)),col=col[1],,ylab="Density",xlab="
   Population",sub="",lwd=2,cex.axis=1.2,cex.lab=1.2,main="")
points(density(DnoRM[,6]),type="l",lty=2,lwd=2,col=col[2])
points(density(DnoRM[,8]),type="l",lty=3,lwd=2,col=col[3])
legend("topright",as.character(c(Year[1],Year[6],Year[8])),col=
   c(col[1],col[2],col[3]),lty=c(1,2,3),lwd=c(2,2,2),cex=1.2,
   bty="n")

quartz()
plot(density(DRM[,1]),xlim=c(0.4,max(M.X.RM)),ylim=c(min(m.Y.RM
   ),max(M.Y.RM)),col=col[1],main="",ylab="Density",xlab="
   Population",sub="",lwd=2,cex.axis=1.2,cex.lab=1.2)
points(density(DRM[,6]),type="l",lty=2,lwd=2,col=col[2])
points(density(DRM[,8]),type="l",lty=3,lwd=2,col=col[3])
legend("topright",as.character(c(Year[1],Year[6],Year[8])),col=
   c(col[1],col[2],col[3]),lty=c(1,2,3),lwd=c(2,2,2),cex=1.2,
   bty="n")

pkg <- "/Users/andresvallone/Dropbox/tesis_AndresVallone/PAPERS
   /Rpackage/03-estdaR/estdaR"

load_all(pkg)

or.d <- apply(b,2,discret)
table(or.d)

full.markov <- mkv(b)
st.st(full.markov[[1]])
```

```
k1 <- knn2nb(knearneigh(coords, k=1))
k1d <- nbdists(k1, coords)
max_k1 <- max(unlist(k1d))
dnn_nb <- dnearneigh(coords, 0, max_k1)
dlist<-nbdists(dnn_nb,coords)
idlist<-lapply(dlist, function(x) 1/x)
Wd<-nb2listw(dnn_nb,glist=idlist,style="W")

# Spatial Markov

spatial.markov <- sp.mkv(b,Wd,fixed=FALSE)

sapply(1:5,function(x)st.st(spatial.markov[[1]][,,x]))

sp.homo.test(b,Wd,fixed=FALSE)


#LISA Markov

L.M <- lisamkv(b,Wd,geoda=FALSE)
st.st(L.M$p.lisamatrix)

join.d(b,Wd)

# Directional LISA

Regime <- ifelse(base$FIELD2=="RM","MR","All the Rest")

#fig 4

full.p8 <- d.LISA(D[,1],D[,8],Wd,Regime=Regime,nsim=999)
full.p4 <- d.LISA(D[,1],D[,8],Wd,Regime=Regime,nsim=999,k=4)

full.p8[[1]]


# fig 5
```

```
# En vez de anlizar el perido inical contra el final, hice un
   pool de todos los moviemntos y lo divido en periodos antes
   de 70 y despues. Se encuentran comportamientos regionales
   distitnos en entre peridos.


# k= 8

step <- (2*pi)/8
breaks <- seq(0,2*pi,step)
lmt <- (breaks * 180)/pi # rad to deg
symb<-c()
for (i in seq_len(length(lmt)-1)){
  symb <- c(symb, paste(lmt[i],lmt[i+1],sep="-"))
}
lim=c(symb[1],symb[length(symb):2])

dL <- lapply(seq_len((ncol(b)-1)),function(x)d.LISA(D[,x],D[,x
   +1],Wd,Regime=Regime,nsim=999))

p1 <- list()
for(i in 1:4){p1[[i]]<-dL[[i]]$data}

p2 <- list()
for(i in 5:7){p2[[i]]<-dL[[i]]$data}

pt <- list()
for(i in 1:7){pt[[i]]<-dL[[i]]$data}


p1 <- do.call(rbind.data.frame,p1)
p2 <- do.call(rbind.data.frame,p2)
pt <- do.call(rbind.data.frame,pt)



# de 1930-1960
```

```
quartz()
rose.p1<-ggplot(p1,aes(angle,fill=Regime),xlab=" ",ylab=" ") +
   geom_bar(width=1,colour="black",size=0.1)+scale_x_discrete(
   name="",limit=lim)+scale_y_continuous(name="",breaks=seq(0,
   max(table(p1$angle)),5),labels=seq(0,max(table(p1$angle)),5)
   )+coord_polar(start=pi/4)+theme(legend.position="bottom",
   panel.grid.major = element_line( color="gray50",linetype="
   solid"),panel.background = element_blank())+scale_fill_
   brewer(palette="Set1",direction=-1)
rose.p1

# 1970-2002
quartz()
rose.p2<-ggplot(p2,aes(angle,fill=Regime),xlab=" ",ylab=" ") +
   geom_bar(width=1,colour="black",size=0.1)+scale_x_discrete(
   name="",limit=lim)+scale_y_continuous(name="",breaks=seq(0,
   max(table(p2$angle)),5),labels=seq(0,max(table(p2$angle)),5)
   )+coord_polar(start=pi/4)+theme(legend.position="bottom",
   panel.grid.major = element_line( color="gray50",linetype="
   solid"),panel.background = element_blank())+scale_fill_
   brewer(palette="Set1",direction=-1)
rose.p2

# total
quartz()
rose.pt<-ggplot(pt,aes(angle,fill=Regime),xlab=" ",ylab=" ") +
   geom_bar(width=1,colour="black",size=0.1)+scale_x_discrete(
   name="",limit=lim)+scale_y_continuous(name="",breaks=seq(0,
   max(table(pt$angle)),5),labels=seq(0,max(table(pt$angle)),5)
   )+coord_polar(start=pi/4)+theme(legend.position="bottom",
   panel.grid.major = element_line( color="gray50",linetype="
   solid"),panel.background = element_blank())+scale_fill_
   brewer(palette="Set1",direction=-1)
rose.pt

# GIMA indicator

# block neighbors matrix
```

```
WR <- matrix(0,nrow(base),nrow(base))
for(i in seq_len(nrow(base))){
  WR[,i] <- 1*(Regime[i]==Regime)
}
diag(WR) <- 0

WR <- mat2listw(WR)

spTau <- lapply(1:(ncol(b)-1),function(i) sp.tau(b[,i],b[,(i+1)
   ],WR,999))

GIMA <- do.call(rbind.data.frame,spTau)

colnames(GIMA)<-names(spTau[[1]])

p <- matrix(1,nrow(b),nrow(b))
diag(p)<-0

phi <- spweights.constants(WR)$S0/sum(p)

nonng.tau <- (GIMA[,5]-(phi*GIMA[,1]))/(1-phi)
ng.tau <- (GIMA[,1])
ng.M <- (ng.tau-1)/-2
nonng.M <- (nonng.tau-1)/-2
pvalue <- GIMA[,2]
per <- c("1930-1940","1940-1952","1952-1960","1960-1970","
   1970-1982","1982-1992","1992-2002")

# tabla 4

tabla.GIMA <- data.frame("Period"=per,"MW"=ng.M,"MnonW"=nonng.M
   ,"TW"=ng.tau,"TnonW"=nonng.tau,"p-value"=pvalue)

# Para robustez uso la matriz de distancia e estimo el TAU que
   da significativo en los mismos periodos.
```

```
spTau2 <- lapply(1:(ncol(b)-2),function(i) sp.tau(D[,i],D[,(i
   +1)],Wd,999))

GIMA2 <- do.call(rbind.data.frame,spTau2)

colnames(GIMA2)<-names(spTau[[1]])

# Rank Decomposition


Reg <- ifelse(base$FIELD2=="RM",1,2)
th <- theta(b,Reg,nsim=999)

# table 5

theta.table <- data.frame("Period"=per,"Theta"=th[,1],"p-value"
   =th[,2])
```

E “estdaR” source code

Algorithm E.1: d.LISA function code

```
#' @import gridExtra RColorBrewer spdep ggplot2 stats graphics
   sp
#'
#'
#' @name d.LISA
#' @rdname d.LISA
#'
#' @title Directional LISA
#' @description Compute the origin standardized movement of a
   spatial unit and its neighbors and the pseudo p-value of the
    directional co-movement.
#'
#' @param x0 a vector containing the variable at initial period
    of analysis
#' @param x1 a vector containing the variable at final period
   of analysis
#' @param W an object of class \code{listw}
#' @param Regime a numeric vector containing the regime to the
   spatial unit belongs
#' @param k a scalar in c(4,8) indicating number of circular
   sectors in rose diagram, by default it is set as 8
#' @param mean.rel a logical, Is the data mean relative?. By
   default it is FALSE
#' @param nsim  number of random spatial permutations for
   calculation of pseudo p-values, the default value is NULL.
#' @param arrow Logical. Do you want to plot the arrows in the
   standardised plot?. If it is set as FALSE the arrow head is
   plot as a point.
```

```
#' @param only numerical indicating the spatial unit to be
   considered in the plot, the NULL value imply the use of all
   the spatial unit in the plot.
#' @details For later...
#'
#' @references Rey, S. J., Murray, A. T., & Anselin, L. (2011).
    Visualizing regional income distribution dynamics. Letters
   in Spatial and Resource Sciences, 4(1), 81  90  .
#'
#' @return a list contaning \describe{
#'     \item{"Lisa"}{A vector scatterplot showing the
   Directinal co-movement of a spatial unit and its neighbors}
#'     \item{"rose"}{A rose diagram of the LISA}
#'     \item{"p.rose"}{pseudo p-values of the direction in the
   rose diagram. Only available when nsimm is not NULL}
#'     \item{"data"}{a data frame contening the data used for
   the lisa scatterplot}
#'     \item{"counts"}{a data frame containing the information
   used for the rose diagram}
#'     \item{"p.value"}{a data frame containing the information
    used in the pseudo p-value graph. Only available when nsim
   is not NULL}
#'  }
#'
#' @examples
#' data(us48)
#' w1queen <- nb2listw(poly2nb(us48))
#' t0<-us48$X1969/mean(us48$X1969)
#' t1 <-us48$X2009/mean(us48$X2009)
#' Regime <-us48$SUB_REGION
#' ok <- d.LISA(t0,t1,w1queen,Regime,k=8,nsim=999)
#'
#' @export



d.LISA <- function(x0,x1,W,Regime=NULL,k=8,mean.rel=FALSE,nsim=
   NULL,arrow=TRUE,only=NULL){
```

```
skip.leg <- ifelse(is.null(Regime)==TRUE,1,0)
if (is.null(Regime)==TRUE) { Regime <- rep(1,length(x0)) }
x0.lag <- lag.listw(W,x0)
x1.lag <- lag.listw(W,x1)
if (mean.rel==TRUE){
  x0 <- x0 - mean(x0)
  x1 <- x1 - mean(x1)
  x0.lag <- x0.lag - mean(x0.lag)
  x1.lag <- x1.lag - mean(x1.lag)
}
Regime <- factor(Regime)
f.point <- as.data.frame(cbind(x1-x0,x1.lag-x0.lag))
y0 <- NULL # unuseful in the code, but avoids a CRAN note.
y1 <- NULL
vtop<-cbind(rep(0,length(x0)),rep(0,length(x0)),f.point,
 Regime)
colnames(vtop)<-c("x0","y0","x1","y1","Regime")
if(!is.null(only)){
  vtop <- vtop[only,]
}
if(arrow==FALSE){
  lisa <-ggplot(vtop)+
    geom_hline(yintercept=0,linetype=1)+
    geom_vline(xintercept=0,linetype=1)+
    geom_point(aes(x=x1, y=y1,color=Regime),size=0.4)+
    theme(
      legend.position="bottom",
      panel.border = element_rect(linetype = "solid", fill =
 NA),
      panel.background = element_rect(fill = NA),
      panel.grid.major = element_line()
    )+xlab("X")+ylab("WX")+scale_color_brewer(palette="Set1",
 direction=-1)
} else {
  lisa <-ggplot(vtop)+
    geom_hline(yintercept=0,linetype=1)+
    geom_vline(xintercept=0,linetype=1)+
```

```
      geom_segment(aes(x=x0, xend=x1, y=y0, yend=y1,color=
   Regime),size=0.4,arrow=arrow(length = unit(0.2, "cm")))+
      theme(
        legend.position="bottom",
        panel.border = element_rect(linetype = "solid", fill =
   NA),
        panel.background = element_rect(fill = NA),
        panel.grid.major = element_line()
      )+xlab("X")+ylab("WX")+scale_color_brewer(palette="Set1",
   direction=-1)
  }


#------------------ROSE---------------------------------------
  rotar <-ifelse(k==8,pi/4,0)
  step <- (2*pi)/k
  breaks <- seq(0,2*pi,step)
  lmt <- (breaks * 180)/pi # rad to deg
  symb<-c()
  for (i in seq_len(length(lmt)-1)){
    symb <- c(symb, paste(lmt[i],lmt[i+1],sep="-"))
  }
  z<-atan2(f.point[,2],f.point[,1]) #get angles
  z <- ifelse(z>0,1*z,(2*pi)+z) #avoid negatives rad
  bins = rep(1, length(z))
    i = 1L
    for(b in breaks){
        bins[z > b] = i
        i = i + 1L
    }
  angle <- factor(bins, levels = seq_along(symb), labels = symb
   )
  if(!is.null(only)){
    angle <- angle[only]
  }
  len <- sp::spDistsN1(as.matrix(vtop[,3:4]),c(0,0))
  real<-hist(z,breaks=breaks,plot=FALSE)$counts
```

```
  d.plot <- data.frame(Regime=factor(vtop$Regime),angle=as.
   character(angle),length=len)
  lim=c(symb[1],symb[length(symb):2])
  rose<-ggplot(d.plot,aes(angle,fill=Regime),xlab=" ",ylab=" ")
    +geom_bar(width=1,colour="black",size=0.1) +scale_x_
   discrete(name="",limit=lim)+scale_y_continuous(name="",
   breaks=seq(0,max(table(angle)),5),labels=seq(0,max(table(
   angle)),5))+coord_polar(start=rotar)+theme(legend.position="
   bottom",panel.grid.major = element_line( color="gray50",
   linetype="solid"),panel.background = element_blank())+scale_
   fill_brewer(palette="Set1",direction=-1)

#__________P VALUES_____________________

  if(!is.null(nsim)){
  id <- seq_along(x0)
  capture <- matrix(0,nsim,k)
  for (m in seq_len(nsim)){
    rid <- sample(id)
    rx0<-x0[rid]
    rx1<-x1[rid]
    rx0.lag <- lag.listw(W,rx0)
    rx1.lag <- lag.listw(W,rx1)
    rf.point <- as.data.frame(cbind(rx1-rx0,rx1.lag-rx0.lag))
    rz<-atan2(rf.point[,2],rf.point[,1])
    rz <- ifelse(rz>0,1*rz,(2*pi)+rz)
    rf<-hist(rz,breaks=breaks,plot=FALSE)$counts
    capture[m,]<-rf
  }
  colnames(capture) <-symb
  larger <- c()
    for (t in seq_along(real)){
      larger <- c(larger, sum(capture[,t]>=real[t]))
    }
  p.l <- nsim - larger
  p <- ifelse(p.l<larger, p.l,larger)
  p.value <-(p+1)/(nsim+1)
  p.plot<-rep(1,length(p.value))
```

```
  p.plot[p.value<=0.001]<-0.001
  p.plot[p.value>0.001 & p.value<=0.01]<-0.01
  p.plot[p.value>0.01 & p.value<=0.05]<-0.05
  #p.value[p.value>0.05 & p.value<=0.1]<-0.1
  p.plot<-factor(p.plot)
  d.pv <- data.frame(type=symb,p.value=p.plot)
  expec <- as.vector(sapply(as.data.frame(capture),mean))
  d.pv <- within(d.pv,type <- factor(type,levels=c(symb[1],symb
   [length(lmt):2])))
  val <- c("0.001" = "darkblue", "0.01" = "steelblue4","0.05"="
   skyblue2", "1"="azure2")
  lab <- c("0.001", "0.01", "0.05", "1")
  are <- which(lab %in% p.plot)
  pv <- ggplot(d.pv,aes(type,fill=p.value))+geom_bar(width=1)+
   coord_polar(start=rotar)+theme(panel.grid.major = element_
   line( color="gray50",linetype="solid"),panel.background =
   element_blank())+ylab("")+xlab("")+scale_y_continuous(breaks
   =NULL)+ scale_fill_manual(values = val,labels = lab ,limits
   = lab)
    }

#________ OUTPUT__________________________

  if(skip.leg==1){
    lisa <- lisa + theme(legend.position="none")
    rose <- rose + theme(legend.position="none")
  } else {
    lisa <- lisa #+ scale_linetype_discrete(name="Regime")
    rose <- rose + theme(legend.position="right")
  }
  if(!is.null(nsim)){
    d.real<- data.frame(type=symb,Counts=real,Expected=expec)
    gridExtra::grid.arrange(lisa,arrangeGrob(rose, pv, nrow =
   2), ncol = 2)
    output<-list(lisamap=lisa,rose=rose,p.rose=pv,data=d.plot,
   counts=d.real,p.value=d.pv)
  }else{
    d.real<- data.frame(type=symb,Counts=real)
```

```
    gridExtra::grid.arrange(lisa,rose,nrow=1,ncol=2)
    output<-list(lisamap=lisa,rose=rose,data=d.plot,counts=d.
   real)
  }
  invisible(output)
}
```

ALGORITHM E.2: discret function code

```
#' @name discret
#' @rdname discret
#'
#' @title Discretization of a continuous variable
#' @description Transform a continous variable into a discret
   variable.
#'
#' @param  x numerical vector
#' @param classes a number of a numeric vector of two or more
   unique cut points giving the number of intervals into which
   x will be cut
#' @param type  an integer between 1 and 9 selecting one of the
    nine quantile algorithms detailed below to be used. For
   more information see the quantile fuction
#' @param ... other argumnt to \code{\link{discret}} function
#'
#' @details for later..
#'
#' @return a vector
#'
#' @examples
#' x <- rnorm(1000)
#' dx <- discret(x,4)
#'
#' @export


discret<-function(x, classes=5,type=7,...){ #subdivide una
   variable en clases
    if(length(classes)!=1){# class definition ad hoc.
```

```
      if(max(classes)==1){ # breaks are probabilities
        aux<-quantile(x,classes,type = type, ...)
        aux<-c(-Inf,aux[2:(length(aux)-1)],Inf)
        output<-as.numeric(cut(x,breaks=aux))
      } else {
        aux<-c(-Inf,classes,Inf)
        output<-as.numeric(cut(x,breaks=aux))
      }
    } else {
      aux <- quantile(x,seq(0,1,1/classes))
      output<-as.numeric(cut(x,breaks=aux,include.lowest=TRUE))
    }
    return(output)
}
```

Algorithm E.3: geary function code

```
#' @name geary
#' @rdname guery
#'
#' @title Mutivariable and univariable Geary's C statistic
#' @description Compute the univariable Anselin(1995) and
   multivarible Anselin(2017) Geary's C statistic
#'
#' @param x a ventor, matrix or data frame containing the
   variables and spatial units
#' @param W a \code{listw} object
#' @param nsim number of random permutations used to the
   compute the pseudo p value. By default it is NULL
#' @param type a character indicating the Geary's C to be
   compute. If it is set as "uni" the standart Geary's C
   statistic is calulated. Set type as "multi" to cumpute the
   multivariable Geary's C.
#' @param nbcom number of comparisons use in the Bonferroni
   multiple comparisons correction. By default it is set aas
   the number of spatial unit to use.
#' @param ... other argument to \code{quad} function. See \code
   {\link{quad}} for more information.
#' @details later...
```

```
#' @return a data frame (more explanation in Details later)
#' @references
#' Anselin, L. (1995). Local Indicators of Spatial
   AssociationLISA. Geographical Analysis, 27(2), 93 115 .
   \url{https://doi.org/10.1111/j.1538-4632.1995.tb00338.x}
#' Anselin, L (2017). A Local Indicator of Multivariate
   Spatial Association: Extending Gearys c. Center for
   Spatial Data Science working papers. University of Chicago.
#'
#' @examples
#' data(Guerry)
#' w1queen<-nb2listw(poly2nb(Guerry))
#' pcc <- cbind(Guerry$pc1,Guerry$pc2)
#' \dontrun{e1 <- geary(Guerry$pc1,w1queen,nsim=999)
#' e2 <- geary(pcc,w1queen,nsim=999)
#' e3 <- geary(pcc,w1queen,type="multi",nsim=999)}
#' @export


geary <- function(x,W,nsim=NULL,type="uni",nbcom=length(W$
   neighbours),...){
  #argum check


  # dim check
  if(is.null(dim(x))==TRUE){
    out <- local.c(x,W=W,nsim=nsim,nbcom=nbcom,...)
  } else {
    switch(type,
           uni={
             y <- lapply(seq_len(ncol(x)), function(i) x[,i])
             names(y)<-colnames(x)
             out <- lapply(y,local.c,W=W,nsim=nsim,nbcom=nbcom
   ,...)
           },
           multi={
             out <- m.local.c(x,W=W,nsim=nsim,nbcom=nbcom)
           }
```

```
    )
  }
  return(out)
}


# Univariate local Geary

local.c<-function(x,W,nbcom=length(W$neighbours),nsim=NULL,...)
   {
   l.c<-rep(0,length(x))
   p.val<-rep(0,length(x))
   p.l.c<-rep(0,length(x))
   m2<-var(x)
   for (i in seq_along(x)){
    neg<-W$neighbours[[i]]
    w<-W$weights[[i]]
    d<-x[i]-x[neg]
    c1<-w*(d^2)
    c1<-(1/m2)*sum(c1)
    l.c[i]<-c1
    if(!is.null(nsim)){
      p<-rep(0,nsim)
      for (j in seq_len(nsim)){
        if (i==1) { aux <- c(x[i],sample(x[2:length(x)])) }
         else if (i==length(x)){
          aux <- c(sample(x[1:(length(x)-1)]),x[i])
        } else{
          aux1 <-sample(c(x[1:(i-1)],x[(i+1):length(x)]))
          aux<- c(aux1[1:(i-1)],x[i],aux1[i:length(aux1)])
          }
        d<-x[i]-aux[neg]
        c2<-w*(d^2)
        c2<-(1/m2)*sum(c2)
        p[j]<-c2
        }
      #plot(density(p))
      #abline(v=c1)
```

```
      above <- p>=c1
      larger <- sum(above)
      lower <- (nsim - larger)
      final <- pmin(larger,lower)
      p.val[i]<- (final+1)/(nsim+1)
      }
    }
  if(!is.null(nsim)){
    p.l.c <-  p.adjust(p.val,method="bonferroni",n=nbcom)
    p.plot <- rep(1,length(p.val))
    p.plot[p.val>=0.01 & p.val<=0.05] <- 0.05
    p.plot[p.val>0.001 & p.val<=0.01] <- 0.01
    p.plot[p.val<=0.001] <- 0.001
    quad.i <- quad(x,W,...)
    cluster <- quad.i*as.integer(p.plot!=1)
    out<-cbind("local.c"=l.c,"p.values"=p.val,"adj. pvalue"=p.l
   .c,"p Map"=p.plot,"Cluster Map"=cluster)
  } else{
    out<-cbind("local.c"=l.c)
    }
  return(out)
}

# Multivariate Local Geary

m.local.c <- function(X,W,nsim=NULL,nbcom=length(W$neighbours))
   {
  if(is.data.frame(X)) X<-as.matrix(X)
  if(dim(X)[2L]<2L) stop("You must provide at least 2 variables
   ")
  lc.x <-apply(X,2,local.c,W=W)
  m.lc <- apply(lc.x,1,mean)
  if(!is.null(nsim)){
    p<-matrix(0,nrow(X),nsim)
    larger <- rep(0,nrow(X))
    id <- seq_len(nrow(X))
    for (t in id){
      for (j in seq_len(nsim)){
```

```
        if (t==1) {
          orden <- c(id[t],sample(id[2L:length(id)]))
        } else if (t==length(id)){
          orden <- c(sample(id[1L:(length(id)-1L)]),id[t])
        } else{
          aux1 <-sample(c(id[1L:(t-1L)],id[(t+1L):length(id)]))
          orden<- c(aux1[1L:(t-1L)],id[t],aux1[t:length(aux1)])
        }
      X1 <- X[orden,]
      rml.c <- c()
      for (k in seq_len(ncol(X))){
        neg<-W$neighbours[[t]]
        w<-W$weights[[t]]
        d<-X1[t,k]-X1[neg,k]
        c1<-w*(d^2)
        c1<-(1/var(X1[,k]))*sum(c1)
        rml.c <- c(rml.c,c1)
        }
      p[t,j]<-mean(rml.c)
      }
    larger[t] <- sum(p[t,]>=m.lc[t])
    lower <- (nsim - larger[t])
    larger[t] <- pmin(larger[t],lower)
    }
    p.value <- (larger+1)/(nsim+1)
    p.adj <- p.adjust(p.value,method="bonferroni",n=nbcom)
    p.plot <- rep(1,length(p.value))
    p.plot[p.value>=0.01 & p.value<=0.05]<-0.05
    p.plot[p.value>0.001 & p.value<=0.01]<-0.01
    p.plot[p.value<=0.001]<-0.001
    cluster <- as.integer(p.plot!=1)
    out<-cbind("multi local c"=m.lc,"p.value"=p.value,"adj
   pvalue"=p.adj,"p Map"=p.plot,"Cluster Map"=cluster)
  } else {
    out<-cbind("multi local c"=m.lc)
  }
  return(out)
}
```

ALGORITHM E.4: Guerry function code

```
#' Data From A.-M. Guerry, "Essay On The Moral Statistics Of
   France"
#'
#' Data extract from \code{\link[Guerry]{Guerry}} R package
#'
#' @docType data
#' @usage data(Guerry)
#'
#' @format A Spatial Object conatianing the original Guerry's
   dataset with two new variables \describe{
#'  \item{pc1}{First principal component of Crimes Against
   Persons, Crimes Against Property, Literacy, Donations,
   Infants Born out of Wedlock, and Suicides data}
#'  \item{pc2}{Second principal component of Crimes Against
   Persons, Crimes Against Property, Literacy, Donations,
   Infants Born out of Wedlock, and Suicides data}
#' }
#'
#' @keywords dataset
#' @examples
#' data(Guerry)
"Guerry"
```

ALGORITHM E.5: homo.test function code

```
#' @name homo.test
#' @rdname homo.test
#'
#' @title Test of time homogeneity
#'
#' @description Computes the Test of time homogeneity  based on
    Bickenbach and Bode (2003)
#'
#' @param x numerical matrix of n spatial unit ans t time
   periods
#' @param classes a number of a numeric vector of two or more
   unique cut points giving the number of intervals into which
   x will be cut
```

```
#' @param pr number of subperiods in which to divide the entire
    sample
#'
#' @details for later.....
#'
#' @return A list coantaning the Q statistic and the LR
   statistic.
#'
#' @references Bickenbach, F. and Bode, E. 2003. Evaluating the
    Markov Property in Studies of Economic Convergence,
   International Regional Science Review, vol. 26, no. 3, 363
    92
#'
#' @examples
#' data(us48)
#' data <- as.data.frame(us48)
#' pci <- data[,10:90]
#' rpci <- pci/matrix(1,dim(pci))%*%colMeans(pci)
#' homo.test(rpci,pr=5)
#'
#' @export


homo.test <- function(x,classes=5,pr=3){
  n <- classes
  p <- mkv(x,classes=classes)[[1]]
  A_i <- rowSums(p>0)
  M <- trunc(ncol(x)/pr)
  s <- c(0,c(1L:pr)*M)
  p_ij <- array(0,dim=c(n,n,pr))
  n_ij <- array(0,dim=c(n,n,pr))
  for( i in 1L:(length(s)-1)){
    sub<-x[,(s[i]+1):s[i+1]]
    aux <- mkv(sub,classes)
    p_ij[,,i] <- aux[[1]]
    n_ij[,,i] <- aux[[2]]
  }
  aux <- mkv(x[,(s[length(s)-1]+1):ncol(x)],classes)
```

```
  p_ij[,,pr] <- aux[[1]]
  n_ij[,,pr] <- aux[[2]]
  b_i <- rowSums(sapply(seq_len(pr),function(x) rowSums(n_ij[,,
   x]))>0)
  p<-array(p,dim=c(n,n,pr))
  q1 <- ((p_ij-p)^2)/(p+1*(p==0))
  q2 <- array(unlist(lapply(seq_len(pr),function(x)diag(rowSums
   (n_ij[,,x]),n,n))),dim=c(n,n,pr))
  q <- sapply(seq_len(pr),function(x) q2[,,x]%*% q1[,,x])
  Q <- sum(q)
  p.s <- (p_ij>0)*(p>0)
  np.s <- 1*(p.s==0)
  l.ratio <- log((p.s*p_ij+np.s)/(p.s*p+np.s))
  LR <- 2*sum(n_ij*l.ratio)
  dof <- sum((A_i-1)*(b_i-1))
  pQ <- 1-pchisq(Q,dof)
  pL.R <- 1-pchisq(LR,dof)
  testQ <- data.frame("Q"=round(Q,2),"p-value"=round(pQ,4),"d.o
   .f" = round(dof,0))
  testLR <- data.frame("LR"=round(LR,2),"p-value"=round(pL.R,4)
   ,"d.o.f"= round(dof,0))
  out <- list("Q"=testQ, "LR"=testLR)
  return(out)
}
```

ALGORITHM E.6: join.d function code

```
#' @name join.d
#' @rdname join.d
#'
#' @title Test the independence in the dynamics of a variable
   and its neighbors
#' @description Performes a Chi square test for test that
   dynamics of a variable is independent of dynamics of its
   neighbors.
#'
#' @param x numerical matrix of n spatial unit ans t time
   periods
#' @param W an objet of listw class.
```

```
#'
#' @details The test decompose the LISA Markov chain in a pair
   of chains, one for the city and other for the neighbors,
   each chain has two states H and L and under the null of
   independence test the co-movement of the chains.
#'
#' @return a list coantaning \itemize{
#'    \item the Chi square statistcas and its p value
#'    \item The LISA markov transition matrix ander the null
#'    }
#'
#' @examples
#' data(us48)
#' data <- as.data.frame(us48)
#' w1queen <- nb2listw(poly2nb(us48))
#' join.d(data[,10:90],w1queen)
#'
#' @export


join.d <- function(x,W){
  t<-dim(x)[2L]
  n<-dim(x)[1L]
  x.mean <- colMeans(x)
  x.bar <- x/x.mean
  x.L<-matrix(0,nrow=n,ncol=t)
    for (i in 1:t){
      x.L[,i]<-lag.listw(W,x[,i])
    }
  x.L.mean <- colMeans(x.L)
  x.L.bar <- x.L/(x.L.mean)
  h.x <- apply(x.bar > 1,2,as.numeric) + 1 # 2 are a h and 1 a
   low
  h.x.l <- apply(x.L.bar > 1,2,as.numeric) + 1
  ################################################
  m.x <- mkv.int(h.x)
  m.x <- m.x/rowSums(m.x)
  m.x.l <- mkv.int(h.x.l)
```

```
  m.x.l <- m.x.l/rowSums(m.x.l)
  A <- matrix(c(1,0,0,0,0,0,1,0,0,0,0,1,0,1,0,0),4,4,byrow=TRUE
   )
  kp <- m.x %x% m.x.l
  aux <- A %*% kp %*% t(A)
  trans <- lisamkv(x,W)$lisamatrix
  t.trans <- rowSums(trans)
  t.h0 <- diag(t.trans) %*% aux
  chi <- m.chi(trans,t.h0)
  output <- list("Chi2"=chi,"Expected"=t.h0)
  return(output)
}
```

ALGORITHM E.7: lisamkv function code

```
#' @import rgdal
#'
#' @name lisamkv
#' @rdname lisamkv
#'
#' @title  Markov for Local Indicators of Spatial Association
#' @description Compute the  Markov transition matrix for Local
    Indicators of Spatial Association
#'
#' @param x numerical matrix of n spatial unit and t time
   periods
#' @param W an objet of listw class.
#' @param ... other argument to \code{quad} function. See \code
   {\link{quad}} for more infomation.
#' @details For later...
#'
#' @return a list cantaning three object
#' \describe{
#'   \item{move}{a data frame indicating which type of LISA
   transition occurred}
#'   \item{lisamatrix}{markov LISA transition matrix}
#'   \item{p.lisamatrix}{markov probability LISA transition
   matrix}
#' }
```

```
#'
#' @examples
#' data(us48)
#' data <- as.data.frame(us48)
#' w1queen <- nb2listw(poly2nb(us48))
#' ll <- lisamkv(data[,10:90],w1queen)
#'
#' @export

lisamkv<-function(x,W,...){
  if(is.null(dim(x))==TRUE) stop("You must provide a matrix
   conteaining n spatial unita and t  time periods")
  t<-dim(x)[2L]
  n<-dim(x)[1L]
  if(t<2L) stop("At least you must provide two time periods")
  LISA <- apply(x,2,quad,W=W,...)
  type <- matrix(c(1:16),nrow=4,ncol=4,byrow=TRUE) #All the
   possible movements
  move<-matrix(0,nrow=n,ncol=(t-1))
  for (i in 1L:(t-1)){
    for (j in 1:n){
    move[j,i] <- type[ LISA[j,i] , LISA[j,i+1] ]
    }
  }
  f.move <- table(factor(move,levels=c(1:16)))
  lmatrix <- matrix(f.move,nrow=4,ncol=4,byrow=TRUE)
  p.lisa<-lmatrix/rowSums(lmatrix)
  output<-list(move,lmatrix,p.lisa)
  names(output)<-c("move","lisamatrix","p.lisamatrix")
  return(output)
}
```

Algorithm E.8: m.chi function code

```
m.chi <- function(tm, tm.h0){
  rs2 <- rowSums(tm.h0)
  rs1 <- rowSums(tm)
  rs2nz <- rs2 > 0
  rs1nz <- rs1 > 0
```

```
  dof1 <- sum(rs1nz)
  dof2 <- sum(rs2nz)
  rs2 <- rs2 + (rs2 == 0)
  dof <- (dof1 - 1) * (dof2 - 1)
  p <- diag(1/rs2) %*% tm.h0
  E <- diag(rs1) %*% p
  num <- (tm - E)^2
  E <- E + (E == 0)
  chi2 <- sum(num/E)
  pvalue <- 1-pchisq(chi2,dof)
  output <-c("Chi2"=chi2, "p-value"=pvalue, "d.o.f"=dof)
  return(output)
  }
```

ALGORITHM E.9: mexico function code

```
#' Mexican states regional income 1940-2000
#'
#' Data from Mexican states regional income from 1940-2000 used
    in Rey and Guiterrez (2010)
#' @docType data
#' @usage data(mexico)
#' @format A data frame contaning...
#' @references Rey, S.J. and M.L. Sastre Gutierrez. (2010) "
   Interregional inequality dynamics in Mexico." Spatial
   Economic Analysis, 5: 277-298.
#' @source \url{https://github.com/pysal/pysal/tree/master/
   pysal/examples/mexico}
#'
#' @keywords dataset
#' @examples
#' data(mexico)
#' str(mexico)
"mexico"
```

ALGORITHM E.10: mkv.int function code

```
# compute the markov transition matrix
# x is a matrix of discret variables
# internal function used in mkv and join.d
```

```
mkv.int <- function(x){
  clases<-as.numeric(unique(as.factor(x)))
  clases<-sort(clases)
  mm<-matrix(0,nrow=length(clases),ncol=length(clases))
    for (i in 1:dim(x)[1]){
      for(j in 1:(dim(x)[2]-1)){
        mm[x[i,j], x[i,j + 1]] <- mm[x[i,j], x[i,j+1]] + 1
      }
    }
    mm
}
```

Algorithm E.11: mkv function code

```
#' @name mkv
#' @rdname mkv
#'
#' @title  Markov transition probability matrix
#' @description Compute thre Markov transition probability
   matrix
#'
#' @param m numerical matrix of n spatial unit ans t time
   periods
#' @param classes a number of a numeric vector of two or more
   unique cut points giving the number of intervals into which
   x will be cut
#' @param fixed logical, if it is TRUE the data are pooled over
    space and time and the quintiles calculated for the pooled
   data
#' @param type an integer between 1 and 9 selecting one of the
   nine quantile algorithms detailed below to be used. For more
    information see the quantile fuction
#' @param ... other argumnt to \code{\link{discret}} function
#' @return a list contaning the markov's trantiiona matrix and
   the markov's transition probability matrix
#'
#' @examples
#' data(us48)
```

```
#' data <- as.data.frame(us48)
#' pci <- data[,10:90]
#' rpci <- pci/matrix(1,dim(pci))%*%colMeans(pci)
#' m<-mkv(rpci)
#'
#' @export

mkv<-function(m,classes=5,fixed=FALSE,type=7,...){
  #Argument checks
  if(is.null(dim(m))==TRUE) stop("You must provide a matrix
   conteaining n spatial unita and t periods of time")
  t<-dim(m)[2]
  n<-dim(m)[1]
  if(t<2) stop("At least you must provide two period of time
   for this analysis")
  # pool or not cuts.
  if (fixed==TRUE){
    if(length(classes)!=1){
      if(max(classes)==1){
        cuts<-quantile(as.vector(m),classes,type = type, ...)
      }
    } else {
      cuts<-quantile(as.vector(m) ,seq(0,1,1/5),type = type,
   ...)
    }
    x<-apply(m,2,discret,classes=cuts)
  } else{
    x<-apply(m,2,discret,classes=classes)
  }
  mm<-mkv.int(x)
  #rownames(mm)<-clases
  #colnames(mm)<-clases
  mm.p <- mm/rowSums(mm)
  output <- list("Probabilities" = mm.p, "Transitions" = mm)
  return(output)
}
```

Algorithm E.12: moran function code

```
#' @name moran
#' @rdname moran
#'
#' @title Univariable and Bivariable Local Moran's I
#' @description Compute the Univariable (Anselin,1995) and
   Bivariable (cite??) Local Moran's I
#'
#'
#' @param x a ventor, matrix or data frame contening the
   variables and spatial units
#' @param W a \code{listw} object
#' @param nsim number of random permutation used to the compute
    the pseudo p value. By default it is NULL
#' @param type a character indicating the local Moran's I to be
    compute. If it is set as "uni" the standart local Moran's I
    statistic is calulated. Set type as "multi" to cumpute the
   bivariable local Moran's I.
#' @param nbcom number of comparisons use in the Bonferroni
   multiple comparisons correction. By default it is set aas
   the number of spatial unit to use.
#' @param ... other argument to \code{quad} function. See \code
   {\link{quad}} for more infomation.
#' @details later...
#' @return a data frame or a list of data frames (more
   explanation in Details later)
#' @references
#' Anselin, L. (1995). Local Indicators of Spatial
   AssociationLISA. Geographical Analysis, 27(2), 93 115 .
   \url{https://doi.org/10.1111/j.1538-4632.1995.tb00338.x}
#'
#' @examples
#' data(Guerry)
#' w1queen<-nb2listw(poly2nb(Guerry))
#' pcc <- cbind(Guerry$pc1,Guerry$pc2)
#'\dontrun{ e4 <- moran(Guerry$pc1,w1queen,nsim=999)
#' e6 <- moran(pcc,w1queen,type="multi",nsim=999,geoda=FALSE)}
#'
```

```
#' @export


moran <- function(x,W,nsim=NULL,type="uni",nbcom=length(W$
   neighbours),...){
  #argum check

  # dim check
  if(is.null(dim(x))==TRUE){
    out <- local.m(x,W=W,nsim=nsim,nbcom=nbcom,...)
  } else {
    y <- lapply(seq_len(ncol(x)), function(i) x[,i])
    names(y)<-colnames(x)
    switch(type,
           uni={
             y <- lapply(seq_len(ncol(x)), function(i) x[,i])
             names(y)<-colnames(x)
             out <- lapply(y,local.m,W=W,nsim=nsim,nbcom=nbcom
   ,...)
           },
           multi={
             pairs <- expand.grid(c(1:ncol(x)),c(ncol(x):1))
             pairs <- pairs[which(pairs[,1]!=pairs[,2]),] #
   avoid the univariable computation.
             out <- list()
             name<-c()
             for( i in seq_len(nrow(pairs))){
               name <- c(name,paste(colnames(x)[pairs[i,1]],
   colnames(x)[pairs[i,2]],sep="-"))
               m.I<-b.local.m(x[,pairs[i,1]],x[,pairs[i,2]],W=W
   ,nsim=nsim,nbcom=nbcom,...)
               out[[i]] <- m.I
             }
             names(out) <- name
           }
    )
  }
  return(out)
```

```
}


# Univariate local Moran

local.m<-function(x,W,nsim=NULL,nbcom=length(W$neighbours),...)
   {
  l.m <- rep(0,length(x))
  p.val <- rep(0,length(x))
  p.l.m <- rep(0,length(x))
  meanx <- mean(x)
  wx <- lag(W,x)
  m.wx <- wx - mean(wx)
  m.x <- x-meanx
  stdx <- sd(x)
  quad <- quad(x,W,...)
  for (i in seq_along(x)){
    neg<-W$neighbours[[i]]
    w<-W$weights[[i]]
    m1 <- w*((x[neg]-meanx)/stdx)
    m1 <- ((x[i]-meanx)/stdx)*sum(m1)
    l.m[i]<-m1
    if(!is.null(nsim)){
    p<-rep(0,nsim)
    for (j in seq_len(nsim)){
      if (i==1) { aux <- c(x[i],sample(x[2:length(x)])) }
        if (i==length(x)){
            aux <- c(sample(x[1:(length(x)-1)]),x[i])
        } else{
            aux1 <-sample(c(x[1:(i-1)],x[(i+1):length(x)]))
            aux<- c(aux1[1:(i-1)],x[i],aux1[i:length(aux1)])
          }
        m2 <- w*((aux[neg]-meanx)/stdx)
        m2 <- ((x[i]-meanx)/stdx)*sum(m2)
        p[j]<-m2
        }
      # plot(density(p))
      # abline(v=m1)
```

```
    above <- p>=m1
    larger <- sum(above)
    lower <- (nsim - larger)
    final <- pmin(larger,lower)
    p.val[i]<- (final+1)/(nsim+1)
    }
  }
  if(!is.null(nsim)){
    p.l.m <-  p.adjust(p.val,method="bonferroni",n=nbcom)
    p.plot <- rep(1,length(p.val))
    p.plot[p.val>=0.01 & p.val<=0.05]<-0.05
    p.plot[p.val>0.001 & p.val<=0.01]<-0.01
    p.plot[p.val<=0.001]<-0.001
    cluster <- quad*as.integer(p.plot!=1)
    out<-cbind("local m"=l.m,"p.value"=p.val,"adj. p.value"=p.l
   .m, "Moran quad"=quad,"p Map"=p.plot,"Cluster Map"=cluster)
    } else {
      out <- cbind("local.m"=l.m,"Moran quad"=quad)
    }
  return(out)
}

#Multivariable local Moran

b.local.m<-function(x,y,W,nsim=NULL,nbcom=length(W$neighbours)
   ,...){
  l.m <- rep(0,length(x))
  p.val <- rep(0,length(x))
  p.l.m <- rep(0,length(x))
  wy <-lag(W,y)
  m.x <- x - mean(x)
  m.wy <- wy - mean(wy)
  zx <- m.x/sd(x)
  zy <- (y-mean(y))/sd(y) #(mean(y) - y)/sd(y) in geoda
  quad <- quad(x,y=y,W=W,...)
  for (i in seq_along(x)){
    neg<-W$neighbours[[i]]
    w<-W$weights[[i]]
```

```
  m1 <- zx[i]*sum(w*zy[neg])
  l.m[i] <- m1
  if(!is.null(nsim)){
      p<-rep(0,nsim)
    for (j in seq_len(nsim)){
        if (i==1) { aux <- c(y[i],sample(y[2:length(y)])) }
        if (i==length(x)){
          aux <- c(sample(y[1:(length(x)-1)]),y[i])
        } else{
          aux1 <-sample(c(y[1:(i-1)],y[(i+1):length(y)]))
          aux<- c(aux1[1:(i-1)],y[i],aux1[i:length(aux1)])
        }
        zaux <-(aux - mean(aux))/sd(aux) # (mean(aux) - aux)/
 sd(aux)
        m2 <- zx[i]*sum(w*zaux[neg])
        p[j]<-m2
    }
  # plot(density(p))
  # abline(v=m1)
    above <- p>=m1
    larger <- sum(above)
    lower <- (nsim - larger)
    final <- pmin(larger,lower)
    p.val[i]<- (final+1)/(nsim+1)
    }
  }
if(!is.null(nsim)){
  p.l.m <-  p.adjust(p.val,method="bonferroni",n=nbcom)
  p.plot <- rep(1,length(p.val))
  p.plot[p.val>=0.01 & p.val<=0.05]<-0.05
  p.plot[p.val>0.001 & p.val<=0.01]<-0.01
  p.plot[p.val<=0.001]<-0.001
  cluster <- quad*as.integer(p.plot!=1)
  out<-cbind("local m"=l.m,"p.value"=p.val,"adj. p.value"=p.l
 .m, "Moran quad"=quad,"p Map"=p.plot,"Cluster Map"=cluster)
} else {
  out<-cbind("local m"=l.m, "Moran quad"=quad)
}
```

```
  return(out)
}
```

ALGORITHM E.13: prais function code

```
#' @name prais
#' @rdname prais
#'
#' @title Prais conditional mobility measure.
#' @description Compute the Prais conditional mobility measure.
#' @param x Markov probability transition matrix.
#'
#' @details Prais' conditional mobility measure for a class is
   defined as:
#'     \eqn{ pr_i = 1 -  p_{i,i}}
#'
#' @return a vector
#'
#' @examples
#' data(us48)
#' data <- as.data.frame(us48)
#' pci <- data[,10:90]
#' rpci <- pci/matrix(1,dim(pci))%*%colMeans(pci)
#' m<-mkv(rpci)
#' prais(m[[1]])
#'
#' @export

prais <- function(x){
  pr <- 1-diag(x)
  return(pr)
}
```

ALGORITHM E.14: quad function code

```
#'
#'@name quad
#'@rdname quad
#'
#'@title Quandrant Allocator
```

```
#'@description Compute the position of a spatial unit and its
   neihbors in the MOran Scatter Plot
#'
#'@param x a vector
#'@param W a \code{listw} object.
#'@param y a vector, use it in case that the lag must be
   compute in another variable.
#'@param geoda logial.  If True use GeoDa quadrant scheme: HH
   =1, LL=2, LH=3, HL=4. If False use PySAL quadrant Scheme: HH
   =1, LH=2, LL=3, HL=4
#'
#'@details explain here what is the geoda
#'@return a vector
#'
#'@examples
#'data(Guerry)
#'w1queen<-nb2listw(poly2nb(Guerry))
#'quad(Guerry$pc1,w1queen)
#'
#'@export
#'

quad <- function(x,W,y,geoda=TRUE){
  if(missing(y)){y <- x}
  if(geoda==TRUE){
    qn <- c("hh"=1,"ll"=2,"hl"=4,"lh"=3)
  } else {
    qn <- c("hh"=1,"ll"=3,"hl"=4,"lh"=2)
  }
  zx <- (x - mean(x))#/sd(x)
  lx <- spdep::lag.listw(W,y)
  zlx <- (lx - mean(lx))#/sd(lx)
  hx <- zx > 0
  hlx <-  zlx > 0
  hh <- hx * hlx
  lh <- (1 - hx) * hlx
  hl <- hx * (1 - hlx)
  ll <- (1 - hx) * (1 - hlx)
```

```
  out <- qn["hh"]*hh + qn["ll"]*ll + qn["hl"]*hl + qn["lh"]*lh
  return(out)
}
```

Algorithm E.15: shorrock function code

```
#' @name shorrock
#' @rdname shorrock
#'
#' @title Shorrock's mobility measure
#' @description Compute the Shorrock's mobility measure
#'
#' @param x Markov probability transition matrix.
#'
#' @return a vector
#'
#' @examples
#' data(us48)
#' data <- as.data.frame(us48)
#' pci <- data[,10:90]
#' rpci <- pci/matrix(1,dim(pci))%*%colMeans(pci)
#' m <- mkv(rpci)
#' shorrock(m[[1]])
#'
#' @export

shorrock <- function(x){
  if(ncol(x)!=nrow(x)) stop(x, "is not a square matrix")
  trace <- sum(diag(x))
  n <- nrow(x)
  sh <- (n - trace)/ (n - 1)
  return(sh)
}
```

Algorithm E.16: sig.lisamkv function code

```
#' @name sig.lisamkv
#' @rdname sig.lisamkv
#'
```

```
#' @title  Markov for Local Indicators of Spatial Association
   including a significant state
#' @description Compute the  Markov transition matrix for Local
    Indicators of Spatial Association based in the significance
    of the local moran indicator
#'
#' @param x numerical matrix of n spatial unit ans t time
   periods
#' @param W an objet of listw class.
#' @param nsim  number of random spatial permutations for
   calculation of pseudo p-values, by default is set in 999
#' @param ... other argument to \code{quad} function. See \code
   {\link{quad}} for more infomation.
#' @details first the the Local Moran indicator is compute
   useing the \code{nsim} number of permutation. The non
   significant indication is considere as an state of the
   markow chain.
#'
#' @return a list cantaning three object
#' \describe{
#'   \item{move}{a data frame indicating which type of LISA
   transition occurred}
#'   \item{lisamatrix}{markov LISA transition matrix}
#'   \item{p.lisamatrix}{markov probability LISA transition
   matrix}
#' }
#'
#' @examples
#' \dontrun{
#' data(us48)
#' data <- as.data.frame(us48)
#' w1queen <- nb2listw(poly2nb(us48))
#' ll <- sig.lisamkv(data[,10:90],w1queen,999,geoda=FALSE)
#'}
#' @export

sig.lisamkv <- function(x,W,nsim=999,...){
```

```
  if(is.null(dim(x))==TRUE) stop("You must provide a matrix
   conteaining n spatial unita and t  time periods")
  t<-dim(x)[2L]
  n<-dim(x)[1L]
  if(t<2L) stop("At least you must provide two time periods")
  l1 <- lapply(1:t,function(i) moran(x[,i],W,nsim=nsim,...))
  states <- matrix(0,n,t)
  for (i in 1:t){
    states[,i] <- l1[[i]][,6L]
  }
  states <- ifelse(states==0,5,states)
  type <- matrix(c(1:25),nrow=5,ncol=5,byrow=TRUE) #All the
   possible movements
  move<-matrix(0,nrow=n,ncol=(t-1))
  for (i in 1L:(t-1)){
    for (j in 1:n){
      move[j,i] <- type[ states[j,i] , states[j,i+1] ]
    }
  }
  f.move <- table(factor(move,levels=c(1:25)))
  lmatrix <- matrix(f.move,nrow=5,ncol=5,byrow=TRUE)
  p.lisa<-lmatrix/rowSums(lmatrix)
  output<-list(move,lmatrix,p.lisa)
  names(output)<-c("move","lisamatrix","p.lisamatrix")
  return(output)
}
```

ALGORITHM E.17: sp.homo.test function code

```
#' @name sp.homo.test
#' @rdname sp.homo.test
#'
#' @title Test for homogeneity of Markov transition
   probabilities across regimes.
#' @description Performs the homogenity across space test for
   spatial markov trasntion matrix basis on Rey et al, (2016)
#'
#' @param x numerical matrix of n spatial unit ans t time
   periods
```

```
#' @param W an objet of listw class.
#' @param classes a number of a numeric vector of two or more
   unique cut points giving the number of intervals into which
   x will be cut
#' @param fixed logical, if it is TRUE the data are pooled over
    space and time and the quintiles calculated for the pooled
   data
#'
#' @details For later...
#'
#' @return A list coantaning the Q statistic, the LR statistic
   and matrix use as null hypotesis in the test.
#'
#' @references S. J. Rey, W. Kang, and L. Wolf (2016)  The
   properties of tests for spatial effects in discrete Markov
   chain models of regional income distribution dynamics,
   Journal of  Geographical Systems, vol. 18, no. 4, pp. 377
     398  .
#'
#' @examples
#' data(us48)
#' data <- as.data.frame(us48)
#' pci <- data[,10:90]
#' rpci <- pci/matrix(1,dim(pci))%*%colMeans(pci)
#' w1queen <- nb2listw(poly2nb(us48))
#' sp.homo.test(rpci,w1queen)
#'
#' @export

sp.homo.test <- function(x,W,classes=5,fixed=TRUE){
  n <- classes
  sm <- sp.mkv(x,W,classes=classes,fixed=fixed)
  M <- sm[[2]]
  B <- matrix(0,n,n)
  T.M <- apply(M,c(1,2),sum)
  tot <- sum(T.M)
  n_i <- rowSums(T.M)
  A_i <- rowSums(T.M>0)
```

```
A_im <- matrix(0,n,n)
p_ij <- diag(1/(n_i+sum(n_i==0)))%*%T.M
den <- p_ij+1*(p_ij==0)
b_i <- A_i*0
p_ijm <- M*0
Q <- 0
L.R <- 0
q.table <- M*0
LR <- M*0
k<-1
for (j in 1:n){
  m <- M[,,j]
  nim<-rowSums(m)
  B[,k] <- 1*(nim>0)
  b_i <- b_i + 1 * (nim > 0)
  p_ijm[,,k] <- diag(1/(nim+sum(nim==0))) %*% m
  num <-  (p_ijm[,,k]-p_ij)^2
  ratio <-  num / den
  qijm <- diag(nim) %*% ratio
  q.table[,,k] <- qijm
  Q <- Q + sum(qijm)
  mask = (m > 0) * (p_ij > 0)
  A_im[, k] = rowSums(m > 0)
  unmask <- 1 * (mask == 0)
  lr.ratio <- (mask * p_ijm[,,k] + unmask) / (mask * p_ij +
 unmask)
  lr = m * log(lr.ratio)
  L.R <- L.R +sum(lr)
  LR[,,k]<-2 * lr
  k+1
}
dof <- as.integer(sum((b_i - 1) * (A_i - 1)))
L.R <- L.R * 2
pQ <- 1-pchisq(Q,dof)
pL.R <- 1-pchisq(L.R,dof)
testQ <- data.frame("Q"=round(Q,2),"p-value"=round(pQ,4),"d.o
 .f" = round(dof,0))
```

```
  testLR <- data.frame("LR"=round(L.R,2),"p-value"=round(pL.R
   ,4),"d.o.f"= round(dof,0))
  out <- list("Q"=testQ, "LR"=testLR, "NULL"=round(p_ij,4))
  return(out)
}
```

ALGORITHM E.18: sp.mkv function code

```
#' @name sp.mkv
#' @rdname sp.mkv
#'
#' @title Spatial Markov transition probability matrix
#' @description Compute the Spatial Markov transition
   probability matrix (Rey,2001)
#'
#' @param m numerical matrix of n spatial unit ans t time
   periods
#' @param W an objet of listw class.
#' @param classes a number of a numeric vector of two or more
   unique cut points giving the number of intervals into which
   x will be cut
#' @param fixed logical, if it is TRUE the data are pooled over
    space and time and the quintiles calculated for the pooled
   data
#'
#' @details for later..
#' @return a list contaning the markov's trantiiona matrix and
   the markov's transition probability matrix
#'
#' @references Rey, S.J. (2001) Spatial empirics for
   economic growth and convergence, 34 Geographical Analysis
   , 33, 195-214.
#'
#' @examples
#' data(us48)
#' data <- as.data.frame(us48)
#' pci <- data[,10:90]
#' rpci <- pci/matrix(1,dim(pci))%*%colMeans(pci)
#' w1queen <- nb2listw(poly2nb(us48))
```

```
#' sp.mkv(rpci,w1queen)
#'
#' @export


sp.mkv <- function(m,W,classes=5,fixed=TRUE){
  lag<-matrix(0,nrow(m),ncol(m))
  for(j in 1:ncol(m)){
    lag[,j]<-lag.listw(W,m[,j])
  }
  if(fixed==TRUE){
    auxl <- unlist(c(lag))
    auxx <- unlist(c(m))
    auxl <- discret(auxl,classes=classes)
    auxx <- discret(auxx,classes=classes)
    lx <- matrix(auxl,dim(lag))
    x <- matrix(auxx,dim(m))
  } else {
    lx <- apply(lag,2,discret,classes=classes)
    x <- apply(m,2,discret,classes=classes)
  }
  class<-as.numeric(unique(as.factor(lx)))
    class<-sort(class)
    mm <- array(0,dim = c(length(class),length(class),length(
   class)),dimnames=list(class,class,paste("Lag",class)))
  for (j in 1:ncol(x)-1){
    for(i in 1:nrow(x)){
             mm[x[i,j], x[i,j+1],lx[i,j]] <- mm[x[i,j], x[i,j
   +1],lx[i,j]] + 1
          }
    }
    mm.p <- array(0,dim = c(length(class),length(class),length(
   class)),dimnames=list(class,class,paste("Lag",class)))
    for(k in seq_along(class)){
        mm.p[,,k] <- mm[,,k]/rowSums(mm[,,k])
    }
  output <- list("Probabilities" = mm.p, "Transitions" = mm)
  return(output)
```

```
}
```

Algorithm E.19: sp.tau function code

```
#' @name sp.tau
#' @rdname sp.tau
#'
#' @title Global Indicators of Mobility Association (GIMA)
#' @description Compute the Global Indicators of Mobility
   Association index Rey(2016)
#'
#' @param x rank variable
#' @param y rank variable.
#' @param W an object of class \code{listw}
#' @param perm  number of random spatial permutations for
   calculation of pseudo p-values, the default value is NULL.
#'
#' @details The Global Indicators of Mobility Association (GIMA
   ) is based on the Spatial Tau indicator (Rey,2004). The
   implementation is a two step algorith based on the Rey's
   implementation (Rey,2014)
#'
#' @return a vector coantaning \itemize{
#'    \item The GIMA statistic value
#'    \item The GIMA's pseudo p value, only available for perm
   >0
#'    \item The number of spatial Concordant pairs
#'    \item The number of spatial Discordant pairs
#'    \item The Tau statistic value
#'    \item The tau's p value
#'    \item The number of Concordant pairs
#'    \item The number of Discordant pairs
#'    \item The number of extra x pairs. An extra x pair is a
   pair which \eqn{sgn(x_i - x_j)=0}
#'    \item The number of extra y pairs. An extra y pair is a
   pair which \eqn{sgn(y_i - y_j)=0}
#'    }
#' @examples
#' data(mexico)
```

```
#' n <- nrow(mexico)
#' w <- matrix(0,n,n)
#' for(i in 1:n){
#'    w[,i] <- 1*(mexico$Regime[i]==mexico$Regime)
#' }
#' diag(w) <- 0
#' W <- mat2listw(w,style="B")
#' a <- lapply(1:(ncol(mexico)-2),function(i) sp.tau(mexico[,i
   ],mexico[,(i+1)],W,999))
#' b<-do.call(rbind.data.frame,a)
#' colnames(b) <- names(a[[1]])
#' b <- round(b,3)
#'
#'
#' @export
```

```
sp.tau <- function(x,y,W,perm=NULL){
  f.step <- tau(x,y)
  Tau <- f.step["Tau"]
  Tau_p <- f.step["pval"]
  concordant <- f.step["Concordant"]
  discordant <- f.step["Discordant"]
  extraX <- f.step["ExtraX"]
  extraY <- f.step["ExtraY"]
  res <- int.tau(x,y,W)
  if(!is.null(perm)){
    taus <- rep(0,perm)
    ids <- seq_along(x)
    for (r in 1:perm){
      rids <- sample(ids)
      taus[r] <- int.tau(x[rids],y[rids],W)[1]
    }
    above <- taus >= res[1]
    larger <- sum(above)
    psim <- (larger + 1)/ (perm + 1)
```

```
    if (psim > 0.5) psim <- (perm - larger +1) / (perm + 1)
    out <- c("sp.tau"=res["tau_g"],"pval"=psim,"sp.Concordant"=
   res["gc"],"sp.Discordant"=res["gd"],f.step)
  } else {
    c("sp.tau"=res["tau_g"],"sp.Concordant"=res["gc"],"sp.
   Discordant"=res["gd"],f.step)
  }
  out
}

int.tau<-function(x,y,W){
  n1 <- n2 <- iS <- gc <- 0
  for(i in seq_along(W$neighbours)){
    neg <- unlist(W$neighbours[i])
    xi <- x[i]
    yi <- y[i]
    for(k in seq_along(neg)){
      j <- neg[k]
      if(i < j){
        xj <- x[j]
        yj <- y[j]
        dx <- xi - xj
        dy <- yi - yj
        dxdy <- dx * dy
        if(dxdy!=0){
          n1 <- n1 + 1
          n2 <- n2 + 1
          if(dxdy > 0){
            gc <- gc + 1
            iS <- iS + 1
          } else {
            iS <- iS - 1
          }
        } else {
          if (dx !=0) n1 <- n1 +1
          if (dy !=0) n2 <- n2 +1
        }
      }
```

```
    }
  }
  tau_g <- iS /(sqrt(n1) * sqrt(n2))
  gd <- gc - iS
  out <- c("tau_g"=tau_g,"gc"=gc,"gd"=gd)
  return(out)
}
```

ALGORITHM E.20: st.st function code

```
#' @name st.st
#' @rdname st.st
#'
#' @title Steady State distribution of a Marov's probability
   transition matrix
#' @description Calulates the steady stae distribution of a
   Marov's probability transition matrix
#'
#' @param m Markov probability transition matrix
#'
#' @details for later...
#'
#' @return a vector coantaning the steady state distribution.
#'
#' @examples
#' t <- matrix(c(.5,.25,.25,.5,0,.5,.25,.25,.5),3,3,byrow=TRUE)
#' st.st(t)
#'
#' @export
#'

st.st <- function(m){
  aux <- eigen(t(m))
  eva <- aux$values
  eve <- aux$vectors
  m.eva <- max(eva)
  i <- which(eva==m.eva)
  return(eve[,i]/sum(eve[,i]))
}
```

ALGORITHM E.21: Tau function code

```
#' @import NORMT3
#' @name tau
#' @rdname tau
#' @title Kendall rank correlation coefficient
#' @description Compute the Kendall's tau rank correlation
   coefficient index
#'
#' @param x rank variable
#' @param y rank variable.
#'
#' @details Kendalls Tau is a non-parametric measure of
   relationships between columns of ranked data. The
   implemtation is based on Rey(2014),
#'
#' @return a vector coantaning \itemize{
#'    \item The Tau statistic value
#'    \item The tau's p value
#'    \item The number of concordant pairs
#'    \item The number of Discordant pairs
#'    \item The number of extra x pairs. An extra x pair is a
   pair which \eqn{sgn(x_i - x_j)=0}
#'    \item The number of extra y pairs. An extra y pair is a
   pair which \eqn{sgn(y_i - y_j)=0}
#'    }
#' @examples
#' a<-c(12,2,1,12,2)
#' b<-c(1,4,7,1,0)
#' cor(a,b,method="kendall")
#' tau(a,b)
#'
#'
#' @export


tau <- function(x,y){
  if (length(x)!=length(y)) stop("The rank variable must have
   the same length")
```

```
n <- length(y)
data <- data.frame("x"=x,"y"=y)
data <- data[order(data$x,data$y),]
perm <- as.numeric(rownames(data))
vals <- y[perm]
ExtraY <- 0
  ExtraX <- 0
  ACount <- 0
  BCount <- 0
  CCount <- 0
  DCount <- 0
  ECount <- 0
  DCount <- 0
  Concordant <- 0
  Discordant <- 0
  #left child's id
  li <- rep(NA,(n-1))
  #right child's id
  ri <- rep(NA,(n-1))
  # number of left descendants for a node
  ld <- rep(0,n)
  # number of values equal to value i
  nequal <- rep(0,n)

  for (i in 2:n){
    NumBefore <- 1
    NumEqual <- 2
    root <- 1
    x0 <- x[perm[(i-1)]]
    y0 <- y[perm[(i-1)]]
    x1 <- x[perm[i]]
    y1 <- y[perm[i]]
    if(x0!=x1){
      DCount <- 0
      ECount <- 1
    } else{
      if(y0 == y1){
        ECount <- ECount +1
```

```
      } else {
        DCount <- DCount + ECount
        ECount <- 1
      }
    }
    root <- 1
    inserting <- TRUE
    while(inserting==TRUE){
      current <- y[perm[i]]
      if(current > y[perm[root]]){
        NumBefore <-  NumBefore + 1 + ld[root] + nequal[root]
        if(is.na(ri[root])){
          ri[root] <- i
          inserting <- FALSE
        } else {
          root <- ri[root]
        }
      } else if(current < y[perm[root]]){
        ld[root] <- ld[root]+1
        if(is.na(li[root])){
          li[root] <- i
          inserting <- FALSE
        } else {
          root <- li[root]
        }
      } else if(current == y[perm[root]]){
        NumBefore <- NumBefore + ld[root]
        NumEqual <- NumEqual + nequal[root] +1
        nequal[root] <- nequal[root] + 1
        inserting <- FALSE
      }

    }
    ACount <- (NumBefore-1) - DCount # substracting 1 for
index different
    BCount <- (NumEqual-1) - ECount # substracting 1 for
index different
```

```
      CCount <- i -(ACount + BCount + DCount + ECount - 1)-1 #
   substracting 1 for index different
      ExtraY <- ExtraY + DCount
      ExtraX <- ExtraX + BCount
      Concordant <- Concordant + ACount
      Discordant <-  Discordant + CCount
    }
    cd <- Concordant + Discordant
    num <- Concordant - Discordant
    tau <- num / sqrt((cd + ExtraX) * (cd + ExtraY))
    v <- (4 * n + 10) / (9 * n * (n - 1))
    z <- tau / sqrt(v)
    pval <- NORMT3::erfc(abs(z/1.4142136))
    out <- c("Tau"=tau,"pval"=pval,"Concordant"=Concordant,"
   Discordant"=Discordant,"ExtraX"=ExtraX,"ExtraY"=ExtraY)
    return(Re(out))
}
```

Algorithm E.22: theta function code

```
#' @name theta
#' @rdname theta
#'
#' @title Rank Depcompostion index
#'
#' @description Compute the Regime mobility measure Rey(2004)
#'
#'
#' @param x a \eqn{nxk} matrix with \eqn{l \geq 2} successive
   columns of a variable are later moments in time
#' @param Regime values corresponding to which regime each
   observation belongs to
#' @param nsim  number of random spatial permutations to
   generate for computationally based inference
#'
```

```
#' @details For sequence of time periods Theta measures the
   extent to which rank changes for a variable measured over n
   locations are in the same direction within mutually
   exclusive and exhaustive partitions (regimes) of the n
   locations.
#'
#' @return A data frame
#'
#' @references Rey, S.J. (2004)  Spatial  dependence in the
   evolution of regional income distributions,  in A. Getis,
   J. Mur and H.Zoeller (eds). Spatial Econometrics and Spatial
    Statistics. Palgrave, London, pp. 194-213.
#'
#' @examples
#' data(mexico)
#' theta(mexico[,1:7],mexico[,8],999)
#'
#' @export

theta <- function(x,Regime,nsim=NULL){
  rank <- apply(as.matrix(x),2,rank)
  thetas <- rep(0,ncol(x)-1)
  for (j in seq_along(thetas)){
    thetas[j] <- theta.int(rank[,j],rank[,j+1],Regime)
  }
  if(!is.null(nsim)){
    rep <- c()
    for (i in seq_len(nsim)){
      r.thetas <- rep(0,ncol(x)-1)
        for (j in seq_along(r.thetas)){
          r.thetas[j] <- theta.int(rank[,j],rank[,j+1],sample(
   Regime))
        }
      rep <- rbind(rep,r.thetas)
      }
    larger <- c()
    for (i in seq_along(thetas)){
      larger[i] <- sum(rep[,i]>thetas[i])
```

```
    }
    lower <- nsim - larger
    final <- pmin(larger,lower)
    p.value<- (final+1)/(nsim+1)
    output <- data.frame("theta"=thetas,"p-value"=p.value)
    } else {
      output <- data.frame("theta"=thetas)
    }
  return(output)
}



theta.int <- function(r0,r1,Regime){
  diff <- r1-r0
  class <- unique(Regime)
  aux <- matrix(0,nrow=length(r0),ncol=length(class))
  for (k in seq_along(class)){
    aux[,k] <- Regime == class[k]
  }
  s.col <- colSums(aux*diff)
  out <- sum(abs(s.col))/sum(abs(diff))
  return(out)
}
```

Algorithm E.23: u48 function code

```
#' Nominal per capita 48 US states time series 1929-2009
#'
#' Data used in spatial_dynamics module of PySAL library
#' @docType data
#' @usage data(us48)
#' @format an Spatial object
#'
#' @source \url{https://github.com/pysal/pysal/tree/master/
   pysal/examples/us_income}
#' @keywords datasets
#'
#'
```

```
"us48"
```

Bibliography

Abud, M. J., Fink, C., Hall, B., & Helmers, C. (2013). The use of intellectual property in Chile, Economic Research Working Paper No. 11. *World Intellectual Property Organization (WIPO)*.

Abud, M. J., Hall, B., & Helmers, C. (2015). An Empirical Analysis of Primary and Secondary Pharmaceutical Patents in Chile. *PLOS ONE*, *10*(4), e0124257.

Anderson, G., & Ge, Y. (2005). The size distribution of Chinese cities. *Regional Science and Urban Economics*, *35*(6), 756–776.

Anselin, L. (1995). Local Indicators of Spatial Association—LISA. *Geographical Analysis*, *27*(2), 93–115.

Anselin, L. (2013). *Spatial econometrics: methods and models*. Spatial econometrics. Springer Science & Business Media.

Anselin, L. (2017). A local indicator of multivariate spatial association: extending geary's c. *Center for Spatial Data Science, University of Chicago*.

Antrop, M. (2004). Landscape change and the urbanization process in Europe. *Landscape and Urban Planning*, *67*(1), 9–26.

Arauzo Carod, J. M. (2005). Determinants of industrial location: An application for Catalan municipalities*. *Papers in Regional Science*, *84*(1), 105–120.

Arauzo-Carod, J.-M., & Viladecans-Marsal, E. (2009). Industrial Location at the Intra-Metropolitan Level: The Role of Agglomeration Economies. *Regional Studies*, *43*(4), 545–558.

Archambault, É. (2002). Methods for using patents in cross-country comparisons. *Scientometrics*, *54*(1), 15–30. doi:10.1023/A:1015654903218

Aumueller, D. (2009). Retrieving metadata for your local scholarly papers.

Baeninger, R. (1997). Redistribución espacial de la población: características y tendencias del caso brasileño. *Notas de población*.

Barroso, W., Quoniam, L., & Pacheco, E. (2009). Patents as technological information in Latin America. *World Patent Information*, *31*(3), 207–215.

Bas, T. G., & Kunc, M. H. (2009). National systems of innovations and natural resources clusters: Evidence from copper mining industry patents. *European Planning Studies*, *17*(12), 1861–1879.

Batty, M. (2017). Geocomputation. *Environment and Planning B: Urban Analytics and City Science*, *44*(4), 595–597.

Beel, J., Langer, S., Genzmehr, M., & Müller, C. (2013). Docear's PDF inspector: title extraction from PDF files. In *Proceedings of the 13th acm/ieee-cs joint conference on digital libraries* (pp. 443–444). ACM.

Benavente, J. M. (2006). The role of research and innovation in promoting productivity in Chile. *Economics of Innovation and New Technology*, *15*(4-5), 301–315.

Bento, A. M., Cropper, M. L., Mobarak, A. M., & Vinha, K. (2005). The Effects of Urban Spatial Structure on Travel Demand in the United States. *The Review of Economics and Statistics*, *87*(3), 466–478.

Berners-Lee, R. F. T., & Masinter, L. (2015). *Uniform Resource Identifier (URI): Generic Syntax, Request for Comments: 3986, January 2005.*

Bhargavan, K., Delignat-Lavaud, A., & Maffeis, S. (2013). Language-based Defenses Against Untrusted Browser Origins. In *Usenix security symposium* (pp. 653–670).

Bhattacharya, S., & Guriev, S. (2006). Patents Vs. Trade Secrets: Knowledge Licensing and Spillover. *Journal of the European Economic Association*, *4*(6), 1112–1147.

Bickenbach, F., & Bode, E. (2003). Evaluating the Markov Property in Studies of Economic Convergence. *International Regional Science Review*, *26*(3), 363–392.

Bivand, R. S. (2014). Geocomputation and open-source software. In *Geocomputation* (p. 329).

Black, D., & Henderson, J. V. (2003). Urban evolution in the USA. *Journal of Economic Geography*, *3*(4), 343–372.

Borgatti, S. P., & Halgin, D. S. (2014). The SAGE Handbook of Social Network Analysis. In *The sage handbook of social network analysis* (pp. 417–433). 1 Oliver's Yard, 55 City Road, London EC1Y 1SP United Kingdom: SAGE Publications Ltd.

Brunsdon, C. (2014). Embedded geocomputation: Publishing text, data and software in a reproducible form. In *Geocomputation* (pp. 416–435). CRC Press.

Brunsdon, C., & Singleton, A. (2015). *Geocomputation: A practical primer*. Sage.

Bustos Validiva, H. (2013). *Historia de Isla de Maipo* (Municipaliad de Isla de Maipo). Municipaliad de Isla de Maipo.

Cambiaso, P. S., Alonso, M. C., & Claro, C. F. (2001). Migraciones internas hacia la Región Metropolitana de Santiago de Chile: una comparación con planteamientos teóricos. *Investigaciones Geográficas*, (35), Pág–1.

Cao, F., Ge, Y., & Wang, J. (2014). Spatial data discretization methods for geocomputation. *International Journal of Applied Earth Observation and Geoinformation*, *26*, 432–440.

Castillo-Fernández, O. (2015). *Web Scraping: Applications and Tools.*

Chang, C.-H., Kayed, M., Girgis, M. R., & Shaalan, K. F. (2006). A Survey of Web Information Extraction Systems. *IEEE Trans. on Knowl. and Data Eng. 18*(10), 1411–1428.

Chen, Z., Wenyin, L., Zhang, F., Li, M., & Zhang, H. (2001). Web mining for Web image retrieval. *Journal of the American Society for Information Science and Technology*, *52*(10), 831–839.

Cheng, T., Haworth, J., & Manley, E. (2012). Advances in geocomputation (1996–2011). *Computers, Environment and Urban Systems*, *36*(6), 481–487.

Couclelis, H. (1998). Geocomputation in context. In *Geocomputation: A primer* (pp. 17–29). John Wiley Chichester.

Crespi, G., & Zuniga, P. (2012). Innovation and Productivity: Evidence from Six Latin American Countries. *World Development*, *40*(2), 273–290.

Da Cunha, J. M. P. (2003). *Urbanización, redistribución espacial de la población y transformaciones socioeconómicas en América Latina*. United Nations Publications.

Da Cunha, J. M. P. (2013). Questions and challenges in studies on Latin-American cities. *Chapters*, 127–152.

Davis, J. C., & Henderson, J. V. (2003). Evidence on the political economy of the urbanization process. *Journal of Urban Economics*, *53*(1), 98–125.

De Mattos, C. A. (1999). Santiago de Chile, globalización y expansión metropolitana: lo que existía sigue existiendo. *EURE (Santiago)*, *25*(76).

Denissen, J. J. A., Neumann, L., & Zalk, M. v. (2010). How the internet is changing the implementation of traditional research methods, people's daily lives, and the way in which developmental scientists conduct research. *International Journal of Behavioral Development*, *34*(6), 564–575.

Desmet, K., & Henderson, J. V. (2015). Handbook of regional and urban economics. In *Handbook of regional and urban economics* (pp. 1457–1517). Elsevier.

Dosi, G., Grazzi, M., & Moschella, D. (2017). What do firms know? What do they produce? A new look at the relationship between patenting profiles and patterns of product diversification. *Small Business Economics*, *48*(2), 413–429.

Dowell, K. G., McAndrews-Hill, M. S., Hill, D. P., Drabkin, H. J., & Blake, J. A. (2009). Integrating text mining into the MGI biocuration workflow. *Database*, *2009*.

Duranton, G. (2016). Determinants of city growth in Colombia. *Papers in Regional Science*, *95*(1), 101–131.

Edelman, B. (2012). Using Internet Data for Economic Research. *Journal of Economic Perspectives*, *26*(2), 189–206.

Eluru, N., Bhat, C. R., Pendyala, R. M., & Konduri, K. C. (2010). A joint flexible econometric model system of household residential location and vehicle fleet composition/usage choices. *Transportation*, *37*(4), 603–626.

Escolano Utrilla, S., & Ortiz Véliz, J. (2004). Cambios de la configuración urbana y «sintaxis del espacio» en ciudades intermedias&58; el caso de La Serena (Chile). *Estudios Geograficos*, *65*(255), 297–320.

Escolano Utrilla, S., Ortiz Véliz, J., & Moreno Mora, R. (2007). Globalización y cambios funcionales recientes en las ciudades del sistema urbano chileno. *Cuadernos Geográficos*, (41).

Fischer, M. M. (2006). *Spatial Analysis and GeoComputation: Selected Essays*. Spatial Analysis and GeoComputation. Berlin Heidelberg: Springer-Verlag.

Fischer, M. M., & Leung, Y. (2001). Geocomputational modelling?techniques and applications: Prologue. In *Geocomputational modelling* (pp. 1–12). Springer.

Foster, I. (2011). How computation changes research. *Switching Codes: Thinking through Digital Technology in the Humanities and the Arts*, 15–37.

Fruchterman, T. M., & Reingold, E. M. (1991). Graph drawing by force-directed placement. *Software: Practice and experience*, *21*(11), 1129–1164.

Furman, J. L., Kyle, M. K., Cockburn, I. M., & Henderson, R. (2006). *Public & private spillovers, location and the productivity of pharmaceutical research.*

Futrelle, R. P., Shao, M., Cieslik, C., & Grimes, A. E. (2003). Extraction, layout analysis and classification of diagrams in PDF documents. In *Document analysis and recognition, 2003. proceedings. seventh international conference on* (pp. 1007–1013). IEEE.

Gahegan, M. (1999). Guest Editorial: What is Geocomputation? *Transactions in GIS*, *3*(3), 203–206.

Geisse, G, & Valdivia, M. (1978). Urbanización e industrialización en Chile. *EURE (Santiago)*.

Geisse, G. (1977). Origen y evolución del sistema urbano nacional. *EURE (Santiago)*, *5*, 37–46.

Getis, A., & Ord, J. K. (1992). The analysis of spatial association by use of distance statistics. *Geographical analysis*, *24*(3), 189–206.

Giuliani, E., Morrison, A., Pietrobelli, C., & Rabellotti, R. (2010). Who are the researchers that are collaborating with industry? An analysis of the wine sectors in Chile, South Africa and Italy. *Research Policy*, *39*(6), 748–761.

Glez-Peña, D., Lourenço, A., López-Fernández, H., Reboiro-Jato, M., & Fdez-Riverola, F. (2014). Web scraping technologies in an API world. *Briefings in bioinformatics*, *15*(5), 788–797.

Gök, A., Waterworth, A., & Shapira, P. (2015). Use of web mining in studying innovation. *Scientometrics*, *102*(1), 653–671.

Graham, M., Hogan, B., Straumann, R. K., & Medhat, A. (2014). Uneven Geographies of User-Generated Information: Patterns of Increasing Informational Poverty. *Annals of the Association of American Geographers*, *104*(4), 746–764.

Grasso, G., Furche, T., & Schallhart, C. (2013). Effective web scraping with OXPath. In *Proceedings of the 22nd international conference on world wide web* (pp. 23–26). ACM.

Gregory, T., & Patuelli, R. (2015). Demographic ageing and the polarization of regions—an exploratory space–time analysis. *Environment and Planning A: Economy and Space*, *47*(5), 1192–1210.

Griffioen, R., de Haan, J., & Willenborg, L. (2014). Collecting clothing data from the Internet. In *Proceedings of meeting of the group of experts on consumer price indexes, may* (pp. 26–28).

Guan, Q., Hu, S., Liu, Y., & Yun, S. (2018). High-performance geocomputation with the parallel raster processing library. In *Geocomputational analysis and modeling of regional systems* (pp. 55–74). Springer.

Guimon, J., Klerkx, L. W. A., & de Saint Pierre, T. (2016). How to bring global R&D into Latin America: Lessons from Chile. *Issues in Science and Technology*, *32*(2), 17–19.

Gutiérrez, M. L. S., & Rey, S. J. (2013). Space-time income distribution dynamics in Mexico. *Annals of GIS*, *19*(3), 195–207.

Hadjar, K, Rigamonti, M, Lalanne, D, & Ingold, R. (2004). Xed: a new tool for extracting hidden structures from electronic documents. In *First international workshop on document image analysis for libraries, 2004. proceedings.* (pp. 212–224).

Henderson, J. V., Shalizi, Z., & Venables, A. J. (2001). Geography and development. *Journal of Economic Geography*, *1*(1), 81–105.

Henríquez, C., Azócar, G., & Romero, H. (2006). Monitoring and modeling the urban growth of two mid-sized Chilean cities. *Habitat International*, *30*(4), 945–964.

Herley, C. (2009). So long, and no thanks for the externalities: the rational rejection of security advice by users. In *Proceedings of the 2009 workshop on new security paradigms workshop* (pp. 133–144). ACM.

Hernandez-Cuevas, C., & Valenzuela, P. D. T. (2004). Strategies to capture biotechnology opportunities in Chile. *Electronic Journal of Biotechnology*, *7*(2), 189–205.

Hooley, T., Wellens, J., & Marriott, J. (2011). *What is Online research?: Using the Internet for social science research.* A&C Black.

Howard, P, Pulcini, C, Levy Hara, G, West, R. M., Gould, I. M., Harbarth, S, & Nathwani, D. (2015). An international cross-sectional survey of antimicrobial stewardship programmes in hospitals. *Journal of Antimicrobial Chemotherapy*, *70*(4), 1245.

Huang, M.-H., Chiang, L.-Y., & Chen, D.-Z. (2003). Constructing a patent citation map using bibliographic coupling: A study of Taiwan's high-tech companies. *Scientometrics*, *58*(3), 489–506.

INAPI. (2016). Chile Estrategia Nacional de Propiedad Industrial.

Instituto Nacional de Estadistical. (2014). Auditoría Técnica a la base de datos del levantamiento censal año 2012.

Instituto Nacional de Estadísticas. (1999). Población de los centros poblados de Chile 1875-1992.

Instituto Nacional de Estadísticas. (2005). Chile: ciudades, pueblos, aldeas y caserío, 2005.

Ioannides, Y., & Overman, H. (2004). Spatial evolution of the US urban system. *Journal of Economic Geography*, *4*(2), 131–156.

Jaffe, A. B., & Trajtenberg, M. (2002). *Patents, citations, and innovations: A window on the knowledge economy*. Patents, citations, and innovations. MIT press.

Jofre-Monseny, J., Marín-López, R., & Viladecans-Marsal, E. (2011). The mechanisms of agglomeration: Evidence from the effect of inter-industry relations on the location of new firms. *Journal of Urban Economics*, *70*(2), 61–74.

Junta de Extremadura. (2017). Atlas Socieconómico de Extremadura - Instituto de Estadística de Extremadura. Retrieved from http://estadistica.gobex.es/web/guest/atlas-socieconomico-de-extremadura

Kahn, M. E., & Schwartz, J. (2008). Urban air pollution progress despite sprawl: The "greening" of the vehicle fleet. *Journal of Urban Economics*, *63*(3), 775–787.

Kendall, M. G. (1962). Rank Correlation Methods.

Krauskopf, M., Krauskopf, E., & Méndez, B. (2007). Low awareness of the link between science and innovation affects public policies in developing countries: The Chilean case. *Scientometrics*, *72*(1), 93–103.

Kumar, S. N. (2015). World towards Advance Web Mining: A Review. *American Journal of Systems and Software*, *3*(2), 44–61.

Lanaspa, L., Pueyo, F., & Sanz, F. (2003). The Evolution of Spanish Urban Structure during the Twentieth Century. *Urban Studies*, *40*(3), 567–580.

Le Gallo, J. (2004). Space-Time Analysis of GDP Disparities among European Regions: A Markov Chains Approach. *International Regional Science Review*, *27*(2), 138–163.

Le Gallo, J., & Chasco, C. (2008). Spatial analysis of urban growth in Spain, 1900–2001. *Empirical Economics*, *34*(1), 59–80.

Lecocq, C., & Van Looy, B. (2009). The impact of collaboration on the technological performance of regions: time invariant or driven by life cycle dynamics? An explorative investigation of European regions in the field of biotechnology. *Scientometrics*, *80*(3), 845–865.

Lederman, D., Messina, J., Pienknagura, S., & Rigolini, J. (2013). *Latin American entrepreneurs: many firms but little innovation*. Latin American entrepreneurs. World Bank Publications.

Leydesdorff, L. (2008). Patent classifications as indicators of intellectual organization. *Journal of the Association for Information Science and Technology*, *59*(10), 1582–1597.

Li, S., Dragicevic, S., Castro, F. A., Sester, M., Winter, S., Coltekin, A., ... Cheng, T. (2016). Geospatial big data handling theory and methods: A review and research challenges. *ISPRS Journal of Photogrammetry and Remote Sensing*, *115*, 119–133.

Liu, Y, & Zhang, M. (2012). Financial websites oriented heuristic anti-phishing research. In *2012 ieee 2nd international conference on cloud computing and intelligence systems* (pp. 614–618).

Longhi, C., Musolesi, A., & Baumont, C. (2014). Modeling structural change in the european metropolitan areas during the process of economic integration. *Economic Modelling*, *37*, 395 –407. doi:https://doi.org/10.1016/j.econmod.2013.10.028

Mage, D., Ozolins, G., Peterson, P., Webster, A., Orthofer, R., Vandeweerd, V., & Gwynne, M. (1996). Urban air pollution in megacities of the world. *Atmospheric Environment*, *30*(5), 681–686.

Marinai, S. (2009). Metadata Extraction from PDF Papers for Digital Library Ingest. In *2009 10th international conference on document analysis and recognition* (pp. 251–255).

Marjanović, M., Bajat, B., Abolmasov, B., & Kovačević, M. (2018). Machine learning and landslide assessment in a gis environment. In *Geocomputational analysis and modeling of regional systems* (pp. 191–213). Springer.

Mehlführer, A. (2009). *Web scraping: A tool evaluation*. na.

Modrego, F., McCann, P., Foster, W. E., & Olfert, M. R. (2015). Regional entrepreneurship and innovation in Chile: a knowledge matching approach. *Small Business Economics*, *44*(3), 685–703.

Morales Valera, M., & Sifontes, D. A. (2014). Las patentes como resultado de la cooperación en l+D en américa latina: Hechos y desafíos. *Investigación y Desarrollo*, *22*(1), 22–38.

National Research Council. (2005). *Expanding Access to Research Data: Reconciling Risks and Opportunities*. Expanding Access to Research Data. National Academies Press.

Nitsch, V. (2001). *City growth in Europe*. Duncker & Humblot.

Nolan, D., & Temple Lang, D. (2014). *XML and Web Technologies for Data Sciences with R*. Springer-Verlag New York.

Nygaard, R. (2015). The use of online prices in the Norwegian Consumer Price Index. In *Meeting of the ottowa group, tokyo, japan*.

Openshaw, S. (2014). GeoComputation. In *Geocomputation* (pp. 1–22). CRC Press.

Openshaw, S., & Abrahart, R. J. (1996). Geocomputation. In *First international conference on geocomputation* (pp. 665–666).

Patel, A., Crooks, A., & Koizumi, N. (2018). Spatial agent-based modeling to explore slum formation dynamics in ahmedabad, india. In *Geocomputational analysis and modeling of regional systems* (pp. 121–141). Springer.

Penman, R. B., Baldwin, T., & Martinez, D. (2009). *Web scraping made simple with sitescraper*. Citeseer.

Pimentel, M. (2000). La reestructuración de los espacios nacionales en los inicios del siglo XXI: continuidad y cambio en la distribución espacial de la población mexicana. *Santiago de Chile, CELADE, mimeo*.

Pinto, P. E., Vallone, A., Honores, G., & González, H. (2017). The dynamics of patentability and collaborativeness in chile: An analysis of patenting activity between 1989 and 2013. *World Patent Information*, *49*, 52 –65. doi:https://doi.org/10.1016/j.wpi.2017.05.004

Polidoro, F., Giannini, R., Conte, R. L., Mosca, S., & Rossetti, F. (2015). Web scraping techniques to collect data on consumer electronics and airfares for Italian HICP compilation. *Statistical Journal of the IAOS*, *31*(2), 165–176.

Popp, D., Juhl, T., & Johnson, D. K. N. (2004). Time In Purgatory: Examining the Grant Lag for U.S. Patent Applications. *Topics in Economic Analysis & Policy*, *4*(1).

Pradhan, B. (2013). A comparative study on the predictive ability of the decision tree, support vector machine and neuro-fuzzy models in landslide susceptibility mapping using gis. *Computers & Geosciences*, *51*, 350–365.

Puertas, O. L., Henríquez, C., & Meza, F. J. (2014). Assessing spatial dynamics of urban growth using an integrated land use model. Application in Santiago Metropolitan Area, 2010–2045. *Land Use Policy*, *38*, 415–425.

Quah, D. T. (1996). Empirics for economic growth and convergence. *European Economic Review*, *40*(6), 1353–1375.

Ramírez, P., Leger, P., & Vallone, A. (2014). Un modelo flexible para la simulación de distribución de ciudades. *Ingeniare. Revista chilena de ingeniería*, *22*(3), 351–362.

Rees, P, & Turton, I. (1998). Guest Editorial. *Environment and Planning A: Economy and Space*, *30*(10), 1835–1838.

Rey, S. (2015). Python Spatial Analysis Library (PySAL): An update and illustration. *Geocomputation: A Practical Primer. London: SAGE*, 233–254.

Rey, S. J. (2001). Spatial empirics for economic growth and convergence. *Geographical Analysis*, *33*(3), 195–214.

Rey, S. J. (2004). Spatial analysis of regional income inequality. *Spatially integrated social science*, *1*, 280–299.

Rey, S. J. (2016). Space–Time Patterns of Rank Concordance: Local Indicators of Mobility Association with Application to Spatial Income Inequality Dynamics. *Annals of the American Association of Geographers*, *106*(4), 788–803.

Rey, S. J. (2018). Code as text: Open source lessons for geospatial research and education. In J.-C. Thill & S. Dragicevic (Eds.), *Geocomputational analysis and modeling of regional systems* (pp. 7–21). doi:10.1007/978-3-319-59511-5_2

Rey, S. J., & Anselin, L. (2010). PySAL: A Python library of spatial analytical methods. *Handbook of applied spatial analysis*, 175–193.

Rey, S. J., & Janikas, M. V. (2006). STARS: Space–Time Analysis of Regional Systems. *Geographical Analysis*, *38*(1), 67–86.

Rey, S. J., Murray, A. T., & Anselin, L. (2011). Visualizing regional income distribution dynamics. *Letters in Spatial and Resource Sciences*, *4*(1), 81–90.

Rey, S. J., Mack, E. A., & Koschinsky, J. (2012). Exploratory Space–Time Analysis of Burglary Patterns. *Journal of Quantitative Criminology*, *28*(3), 509–531.

Rey, S. J., Kang, W., & Wolf, L. (2016). The properties of tests for spatial effects in discrete Markov chain models of regional income distribution dynamics. *Journal of Geographical Systems*, *18*(4), 377–398.

Rodriguez, J. (2007). Paradojas y contrapuntos de dinámica demográfica metropolitana : Algunas respuestas basadas en la explotación intensiva de microdatos censales. In *Santiago de Chile : Movilidad espacial y reconfiguración metropolitana, de Mattos, C and Hidalgo, R eds* (pp. 19–52). 8. EURE-Libros.

Rodríguez, J. (2007). United Nations expert group meeting on population distribution, urbanization, internal migration and development.

Rodríguez, J., & Da Cunha, J. M. P. (2009). Urban growth and mobility in Latin America. *Demographic Transformations and Inequalities in Latin America. Historical trends and recent patterns*.

Rodríguez, J., González, D., Ojeda, M., Jiménez, M., & Stang, F. (2009). El sistema de ciudades chileno en la segunda mitad del siglo XX: entre la suburbanización y la desconcentración (The Chilean City System during the Second Half of the 20th Century: Between Sub-Urbanization and Decentralization). *Estudios Demográficos y Urbanos*, *24*(1 (70)), 7–48.

Rogan, J., Franklin, J., Stow, D., Miller, J., Woodcock, C., & Roberts, D. (2008). Mapping land-cover modifications over large areas: A comparison of machine learning algorithms. *Remote Sensing of Environment*, *112*(5), 2272–2283.

Salamone, S., Scannapieco, S. M., & Scarnò, M. (2014). Web scraping and web mining: new tools for official statistics. *Proceedings of Societa Italiana di Statistica (SIS 2014), Cagliari, Sardegna*.

Santiago, C. M., Raggi, J. P. F., & Erices, L. V. (2016). Urban growth trends in midsize Chilean cities: the case of Temuco. *urbe. Revista Brasileira de GestÃUrbana*, *8*, 375–389.

Sayas, J. P. (2006). Urban sprawl in the periurban coastal zones of Athens. *Eπßotaetaέmegaρηση Koßotανmegaνßotακήmegaν Ερενήmegaν*, *121*(121), 71–104.

Schmal, R., López, M. d. S., & Cabrales, F. (2006). El camino hacia la patentación en las Universidades. *Ingeniare. Revista chilena de ingeniería*, *14*(3), 172–186.

Scott, J. (2013). *Social network analysis*. Sage.

Siewert, W., & Udani, A. (2016). Missouri Municipal Ethics Survey: Do Ethics Measures Work at the Municipal Level? *Public Integrity*, *18*(3), 269–289.

Simon, M., Christian, R., Peter, M., & Dominic, N. (2015). *Automated Data Collection with R: A Practical Guide to Web Scraping and Text Mining*.

Soto, J., & Paredes, D. (2016). Cities, wages, and the urban hierarchy. *Journal of Regional Science*, *56*(4), 596–614.

Stallman, R. M. (2002). What is free software. *Free Society: Selected Essays of*, *23*.

Tang, W., Feng, W., Deng, J., Jia, M., & Zuo, H. (2018). Parallel computing for geocomputational modeling. In *Geocomputational analysis and modeling of regional systems* (pp. 37–54). Springer.

Thaiprayoon, S., & Haruechaiyasak, A. K. C. (2016). PDF Extraction Based on Lexical Analysis for Thai Texts. *International Journal of Applied Computer Technology and Information Systems*, *5*(1).

Thakuriah, P. V., Tilahun, N. Y., & Zellner, M. (2017). Big data and urban informatics: Innovations and challenges to urban planning and knowledge discovery. In *Seeing cities through big data* (pp. 11–45). Springer.

Thill, J.-C., & Dragicevic, S. (2018). GeoComputational Analysis and Modeling of Regional Systems. In *Geocomputational analysis and modeling of regional systems* (pp. 3–6). Springer.

Vallone, A. (2018). msp: R package: function to interactively harmonize data with accuracy problems.

Vallone, A., Chasco, C., & Sanchez, B. (2017). DataSpa: Functions to collect Spanish data at municipality level.

Vallone, A., Le Gallo, J., Chasco, C., & Ayouba, K. (2018). estdaR: functions to perfomrs exploratory spatio-temporal data analysis in R.

van Zeebroeck, N., de la Potterie, B. v. P., & Han, W. (2006). Issues in measuring the degree of technological specialisation with patent data. *Scientometrics*, *66*(3), 481–492.

Vaz, E. (2016). The future of landscapes and habitats: The regional science contribution to the understanding of geographical space. *Habitat International*, *51*, 70–78.

Vaz, E. (2018). Regional intelligence: A new kind of giscience. *Habitat International*, *72*, 1 –2. Regional Intelligence: A new kind of GIScience. doi:https://doi.org/10.1016/j.habitatint.2017.11.015

Venables, A. J. (2005). Spatial disparities in developing countries: cities, regions, and international trade. *Journal of Economic Geography*, *5*(1), 3–21.

Wang, H., Fu, L., Lin, X., Zhou, Y., & Chen, J. (2009). A bottom-up methodology to estimate vehicle emissions for the Beijing urban area. *Science of The Total Environment*, *407*(6), 1947–1953.

Wang, H. (2018). Incorporating urban spatial structure in agent-based urban simulations. In *Geocomputational analysis and modeling of regional systems* (pp. 143–165). Springer.

Wickham, H. (2016). Package 'rvest'. *URL: https://cran. r-project. org/web/packages/rvest/rvest. pdf*.

Winkler, W. E. (1990). String Comparator Metrics and Enhanced Decision Rules in the Fellegi-Sunter Model of Record Linkage.

Wright, K. B. (2005). Researching Internet-Based Populations: Advantages and Disadvantages of Online Survey Research, Online Questionnaire Authoring Software Packages, and Web Survey Services. *Journal of Computer-Mediated Communication*, *10*(3), 00–00.

Xiufang, R., Zhongwu, Z., Yajie, S., Taotao, G., Donghua, W., Geography, S. o., & University, Shanxi Normal. (2015). Temporal and Spatial Evolution of the Cities along China Section of the. *Journal of Desert Research*, (1), 248–252.

Xu, Z., & Zhu, N. (2009). City Size Distribution in China: Are Large Cities Dominant? *Urban Studies*, *46*(10), 2159–2185.

Zuhair, H., Selamat, A., & Salleh, M. (2016). New Hybrid Features for Phish Website Prediction. *International Journal of Advances in Soft Computing & Its Applications*, *8*(1).